Working with Scarce Annotations in Computational Pathology

by

Navid Alemi Koohbanani

Thesis

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy

Department of Computer Science

November 2020

Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my deepest gratitude to my PhD supervisor Prof. Nasir Rajpoot who has supported me throughout my PhD journey. His patience, enthusiasm, immense knowledge and hard-working attitude inspired me in my research path. I learned a lot from him and his mentorship made me a capable researcher. I also would like to thank my PhD advisors, Dr. Abhir Bhalerao, and Dr. Till Bretschneider for their constructive feedback in annual review meetings. I also owe a very important debt to Dr. Ali Gooya for his valuable feedbacks and ideas on the first chapters of this work. I am grateful to my internal and external examiners, Dr. Fayyaz Minhaz and Dr. Lee Cooper for their valuable time and insightful comments on my thesis.

I am very grateful to my clinical collaborators: Dr. Ksenija Benet, Dr. Ali Khurram, Dr. Ayesha Azam and Dr. David Snead for their assistance in collecting annotation and teaching me some aspect of pathology.

I received generous support from Dr. Pavitra Krishnaswamy during my stay in Singapore. She was a great advisor and I would like to thank her for her valuable suggestions on my research. I am indebted to lab members in Singapore for their support and assistance: Balagopal Unnikrishnan , Russ Chua, Francisco Leitao and Dr. Leonit Zeynalvand.

I have greatly benefited from current and previous lab members: Dr. Talha Qaiser, Dr. Muhammad Shaban, Jevgenij Gamper, Dr. Simon Graham, Dr. Najah Alsubaie, Dr. Tzu-Hsi Song, Dr. Mary Shapcott, Ruqayya Awan, Saad Bashir, Dr. Nima Hatami, Dr. Mohsin Bilal, Dr. Moazam Faraz, Dr. Shan Raza, Dr. Fayyaz Minhas, Dr. Sajid Javid, Hammam Alghamdi, John Pocock, Sirijay Deshpande and Rawan Albusayli.

Discussions with Shaban, Talha and Jev have been really enjoyable and I appreciate their support and insightful comments on my research. The

gatherings and fun moments that was organized by Hammam are unforgettable. I would like to thank my old friend, Mostafa, who have collaborated with me on several projects. Collaborating with him was a good experience which led to several nice ideas and publications.

I thank my family, particularly my parents, Ali and Badri, whom I owe everything I have, for their efforts, and kind supports during my educational life. My mother prayers and her guidance were always helpful in putting me on the right way.

I dedicate this thesis to my lovely wife, Nadia, who supports me through thick and thin. Thank you for your kindness and grace.

Sponsorships and Grants

I would like to acknowledge the financial support by Intel via the The Alan Turing Institute's strategic partnership with Intel.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. I declare that, except where acknowledged, the material presented in this thesis is my own work, and has not been previously submitted for obtaining an academic degree.

Navid Alemi Koohbanani

November 2020

Publications

First-Authored Publications

Journal Articles

- **Navid Alemi Koohbanani**, Mostafa Jahanifar and Nasir Rajpoot: NuClick: "A Deep Learning Framework for Interactive Segmentation of Microscopic Images." Medical Image Analysis, 65, 10 2020 (2020).

- **Navid Alemi Koohbanani**, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot: "Self-Path: Self-supervision for Classification of Pathology Images with Limited Annotations." Pending Minor revision in Transactions on Medical Imaging (2020)

Conference and Workshop Papers

- **Navid Alemi Koohbanani** , Mostafa Jahanifar, Ali Gooya, and Nasir Rajpoot. "Nuclear instance segmentation using a proposal-free spatially aware deep learning framework." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 622-630. Springer, Cham, 2019.

- **Navid Alemi Koohbanani**,Jevgenij Gamper, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. "PanNuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification." In European Congress on Digital Pathology, pp. 11-19. Springer, Cham, 2019.

- **Navid Alemi Koohbanani**, Mostafa Jahanifar, Ali Gooya, and Nasir Rajpoot. "Nuclei detection using mixture density networks." In International Workshop on Machine Learning in Medical Imaging, pp. 241-248. Springer, Cham, 2018.

- **Navid Alemi Koohbanani**, Awan, Ruqayya, Muhammad Shaban, Anna Lisowska, and Nasir Rajpoot. "Context-aware learning using transferable features for classification of breast cancer histology images."

In International Conference Image Analysis and Recognition, pp. 788-795. Springer, Cham, 2018.

- **Navid Alemi Koohbanani**, Talha Qaisar, Muhammad Shaban, Jevgenij Gamper, and Nasir Rajpoot. "Significance of hyperparameter optimization for metastasis detection in breast histology images." In Computational Pathology and Ophthalmic Medical Image Analysis, pp. 139-147. Springer, Cham, 2018.

Co-Authored Publications

Journal Articles

- Sajid Javed, Arif Mahmood, Muhammad Moazam Fraz, **Navid Alemi Koohbanani**, Ksenija Benes, Yee-Wah Tsang, Katherine Hewitt, David Epstein, David Snead, and Nasir Rajpoot. "Cellular community detection for tissue phenotyping in colorectal cancer histology images." Medical Image Analysis (2020): 101696.

- Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng Ann Heng, Jiahui Li, Zhiqiang Hu, Yunzhi Wang, **Navid Alemi Koohbanani**, Mostafa Jahanifar, Neda Za- mani Tajeddin, Ali Gooya, Nasir Rajpoot, Xuhua Ren, Sihang Zhou, Qian Wang, Dinggang Shen, Cheng Kun Yang, Chi Hung Weng, Wei Hsiang Yu, Chao Yuan Yeh, Shuang Yang, Shuoyu Xu, Pak Hei Yeung, Peng Sun, Amir- reza Mahbod, Gerald Schaefer, Isabella Ellinger, Rupert Ecker, Orjan Smedby, Chunliang Wang, Benjamin Chidester, That Vinh Ton, Minh-Triet Tran, Jian Ma, Minh N Do, Simon Graham, Quoc Dang Vu, Jin Tae Kwak, Akshayku- mar Gunda, Raviteja Chunduri, Corey Hu, Xiaoyang Zhou, Dariush Lot , Reza Safdari, Antanas Kascenas, Alison ONeil, Dennis Eschweiler, Johannes Stegmaier, Yanping Cui, Baocai Yin, Kailin Chen, Xinmei Tian, Philipp Gru- ening, Erhardt Barth, Elad Arbel, Itay Remer, Amir Ben-Dor, Ekaterina Sirazitdinova, Matthias Kohl, Stefan Braunewell, Yuexiang Li, Xinpeng Xie, Linlin Shen, Jun Ma, Krishanu Das Baksi, Mohammad Azam Khan, Jaegul Choo, Adri n Colomer, Valery Naranjo, Linmin Pei, Khan M Iftekharuddin, Kaushiki Roy, Debotosh Bhattacharjee, Anibal Pedraza, Maria Gloria Bueno, Sabari- nathan Devanathan, Saravanan Radhakrishnan, Praveen Koduganty, Zihan Wu, Guanyu Cai, Xiaojie Liu, Yuqin Wang, Amit Sethi. "A Multi-Organ Nu- cleus Segmentation Challenge." IEEE Transactions on Medical Imaging (2019).

- Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat To,

Muhammad Shaban, Talha Qaiser, **Navid Alemi Koohbanani**, Syed Ali Khur- ram, Jayashree Kalpathy-Cramer, Tianhao Zhao, Rajarsi Gupta, Jin Tae Kwak, Nasir Rajpoot, Joel Saltz, Keyvan Farahani. "Methods for Segmen- tation and Classi cation of Digital Microscopy Tissue Images." Frontiers in Bioengineering and Biotechnology, 7 (2019).

Conference and Workshop Papers

- Simon Graham, Muhammad Shaban, Talha Qaiser, **Navid Alemi Koohbanani**, Syed Ali Khurram, and Nasir Rajpoot. "Classification of lung cancer histology images using patch-level summary statistics." In Medical Imaging 2018: Digital Pathology, vol. 10581, p. 1058119. International Society for Optics and Photonics, 2018.

- Yanning Zhou, Simon Graham, **Navid Alemi Koohbanani**, Muhammad Shaban, Pheng-Ann Heng, and Nasir Rajpoot. "Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images." In Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 0-0. 2019.

Abstract

Computational pathology is the study of algorithms and approaches that facilitate the process of diagnosis and prognosis of primarily from digital pathology. The automated methods presented in computational pathology decrease the inter and intra-observability in diagnosis and make the workflow of pathologists more efficient. Digital slide scanners have enabled the digitization of tissue slides and generating whole slide images (WSIs), allowing them to be viewed on a computer screen rather than through a microscope. Digital pathology images present an opportunity for development of new algorithms to automatically analyse the tissue characteristics.

In this thesis, we first focus on the development of automated approaches for detection and segmentation of nuclei. In this regard, for nuclear detection, each nucleus is considered as a Gaussian shape where the mean of Gaussian determines the centroids of nuclei. We investigate the application of mixture density networks for detection of nuclei in the histology images.

We also propose a convolutional neural network (CNN) for instance segmentation of nuclei. The CNN uses the nuclei spatial information as the target to separate the clustered nuclei. Pixels of each nucleus are replaced with the spatial information of that nucleus. The CNN also utilises dense blocks to reduce number of parameters and positional information at different layer of the network to better learn the spatial information embedded in ground truth.

Two chapters of this thesis are dedicated to dealing with lack of annotations in computational pathology. To this end, we propose a method named as NuClick to generate high quality segmentations for glands and nuclei. NuClick is an interactive CNN based method, that requires minimum user interaction for collecting annotations. We show that one click inside a nucleus can be

enough to delineate its boundaries. Moreover, for glands that are more complex and larger objects a squiggle can extract their precise outline.

In another chapter, we propose Self-Path, a method for semi-supervised learning and domain alignment. The main contribution of this chapter is proposing self-supervised tasks that are specific to histology domain and can be extremely helpful when there are not enough annotations for training deep models. One of these self-supervised tasks is predicting the magnification puzzle which is the first domain specific self-supervised task shown to be helpful for domain alignment and semi-supervised learning for classification of histology images.

Nuclear localization allows further exploration of digital biomarkers and can serve as a fundamental route to predicting patient outcome. In chapter 6, by focusing on the challenge of weak labels for whole slide images (WSIs) and also utilising the nuclear localisation techniques, we explore the morphological features from patches that are selected by the model and we observe that these features are associated with patient survival.

Acronyms

AUC Area Under the Curve

C-Index Concordance Index

CI Confidence Interval

CNN Convolutional Neural Network

Cpath Computational Pathology

DQ Detection Quality

DSS Disease Specific Survival

GT Ground Truth

HR Hazard Ratio

IHC Immunohistochemical

KM Kaplan-Meier

LUAD Lung Adenocarcinoma

LUSC Lung Squamous Cell Carcinoma

MDN Mixture Density Network

MIL Multiple Instance Learning

MSE Mean Squred Error

NSCLC Non-Small Cell Lung Carcinoma

OSCC Oral Squamaous Cell Carcinoma

PQ Panoptic Quality

SQ Segmentation Quality

TCGA The Cancer Genome Atlas

UHCW University Hospitals Coventry and Warwickshire

WSI Whole Slide Image

Chapter 1

Introduction

1.1 Cancer

Cancer is a term for a large group of diseases caused when abnormal cells divide rapidly and uncontrollably and can then spread to other tissue and organs (cancer metastasis). In human body millions of cells divide and grow to replace old or dead cells. When cancer develops, cells become more and more abnormal and they do not die, new cells form and grow without stopping which may form masses called tumours. Cancer is the second leading cause of death globally [1]. There were roughly 18 million cancer cases around the world in 2018, of these 9.5 million cases were in men and 8.5 million in women. Lung and breast cancers are the most common cancers worldwide Lung. prostate, colorectal, stomach and liver cancer are the most common types of cancer in men, while breast, colorectal, lung, cervical and thyroid cancer are the most common among women [2]. If cancer is diagnosed at the earliest stage, there would be the highest chance for a cure. Cancer severity and progression is mainly determined by stage and grade. The cancer grade indicates the level of abnormality of cells and tissue when they are viewed under microscope, whereas cancer's stage explains how large the primary tumor is and how far the cancer has spread in the patient's body [3].

As is the case with other medical conditions, there are many signs and symptoms that may indicate the presence of cancer. These may be observed directly, through imaging technologies, or confirmed by lab tests. A biopsy (removal of tissue for microscopic evaluation) is preferred to establish or rule out a diagnosis of cancer. Tissue samples can be easily retrieved from a tumor near the body's surface. If the mass is inaccessible, an imaging exam that enables a tumor to be located precisely and visualized may be ordered before the biopsy is performed [4].

The histological type is determined by microscopic examination of suspected tissue that has been excised by biopsy or surgical resection. If the histological

type is different from what is usually found in the tissue being examined, it can mean the cancer has spread to that area from some primary site. Metastasis can occur by direct extension, via the blood stream or the lymphatic system, or by seeding or implantation of cancer cells [5].

After obtaining the tissue specimen, it is preserved by freezing or paraffin embedding. The frozen or paraffin embedded tissue blocked are then sliced to the section of 3-5μm thickness using microtome (a tool to slice tissue blocks to thin slices). Tissue slices are then mounted on glass slide. These tissue slides are colorless and their components are not distinguishable under microscope. Therefore, they are stained with special chemical markers to highlight different tissue structures. Hematoxylin and Eosin (H&E) are the most common stains which are used in routine pathology practices. Hematoxylin binds with nucleic acids (DNA, RNA) and dyes dense nuclei as dark blue or violet, whereas Eosin dyes cytoplasmic substance as pink, including proteins, nutrients and muscles (connective) tissues. Immunohistochemistry (IHC) is another staining protocol that uses specific antibodies to highlight the presence of hormone receptors such as estrogen (ER), progestrone (PR) and human epidermal growth factor receptor. After staining, the tissue slides are visually examined by an expert pathologist under the optical microscope to determine if the specimen contains any sort of abnormality or malignancy.

1.2 Histological Analysis

Visual examination of morphological features, quantifying the density of tumour rich areas, analysing spatial arrangement and structure of tumour cells, under the microscope or via digital images in histological sections of a tissue, is the basis for disease diagnosis or disease prognosis. After determining cancer grade or stage, the essential treatment options can be selected. Range of cancers are studied in this thesis, but two of the most extensively studied types are Lung cancer and Oral Squamous Cell Carcinoma. In the next two subsections, we provide general overview of these cancer types and provide some common histological characteristics.

1.2.1 Oral Squamous Cell Carcinoma (OSCC)

Oral cancer is where a tumour develops in a part of the mouth. It may be on the surface of the tongue, the inside of the cheeks, the roof of the mouth (palate), the lips or gums, in the glands that produce saliva, the tonsils at the back of the mouth, and the part of the throat connecting the mouth to windpipe (pharynx). Over 350,000 people worldwide will be diagnosed with oral cancer this year. It will cause over 170,000 deaths, killing roughly one

person every 3 minutes [2].

This cancer is categorised by the type of cancerous cells where squamous cell carcinoma (SCC) is the most common type of mouth cancer, accounting for 9 out of 10 cases. SCC is the cancer starting in the squamous cells. Squamous cells are the flat, skin like cells covering the inside of the mouth, nose, larynx and throat. It has a significant recurrence rate and frequently metastasizes to cervical lymph nodes.

1.2.1.1 OSCC metastasis to lymph node

In patients diagnosed with tumors at an advanced stage, there is a high probability of metastasis to surrounding tissues, particularly cervical lymph node and distant metastasis. The OSCC has high risk of second malignancy during the patient's lifetime. Lymph node metastatic tumors occur in about 40% of patients with oral cancer [6]. The presence of neck lymph node metastasis (NLNM) is universally accepted as the main factor reducing survival in patients with squamous cell carcinoma (SCC) of the oral and oropharyngeal mucosa [7]. An example of histolgy image of SCC metastasis to the lymph node is shown in figure 1.1.

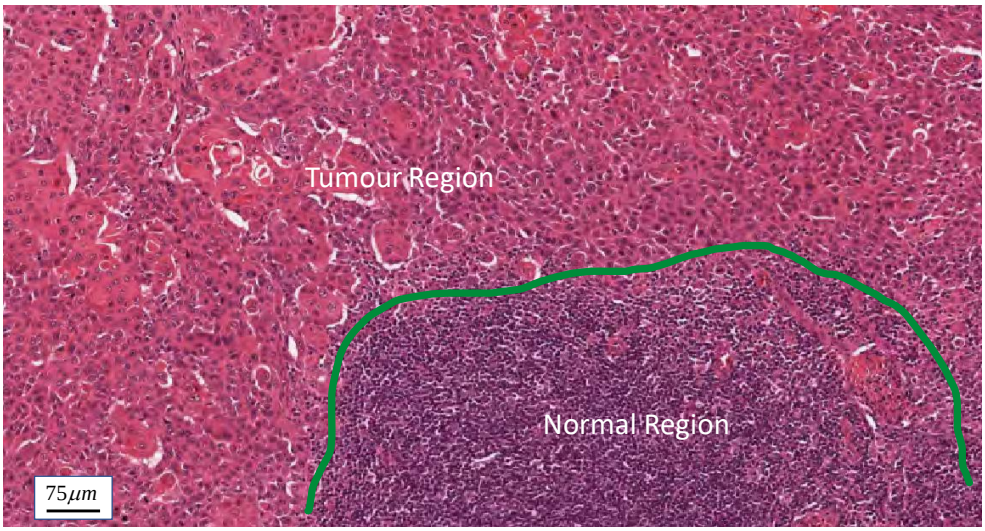

Figure 1.1: Image region from cervical lymph node indicating the normal and oral cancer metastasis region.

1.2.2 Lung Cancer

Lung Cancer is the most common type of cancer worldwide. It remains one of the leading causes of death in several countries including the UK. It is the 3rd most common cancer in the UK, accounting for 13% of all new cancer cases (2017). There are two types of lung cancer: Non-small cell cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC accounts for nearly 90% of lung

cancer diagnoses and develops in the slower rate compared to SCLC. There are three main types of NSCLC: Lung Adenocarcinoma (LUAD), Lung Squamous cell (LUSC) and Large-cell undifferentiated carcinoma. The main histological differences between these categories are cell origin and the morphology of epithelial tumour cell. Adenocarcinoma are found in the peripheral part of lung and in glands that secret mucus. Presence of glandular structure in the tissue can indicate the adenocarcinoma (adeno means gland). Squamous cell cancer is found in the central part of lung. Squamous cell lung cancer accounts for around 30% of all non-small cell lung cancers and is commonly associated with smoking. Large-cell undifferentiated carcinoma can be found anywhere in the lung. This type of cancer grows and spreads very quickly [8].

Adenocarcinoma is histologically categorized into one of the following histology growth patterns: lepidic, acinar, Papillary, micropapillary, solid and invasive mucinous. Figure 1.2 shows different patterns in the LUAD. Due the the heterogeneity of tumour in LUAD, it is not trivial to distinguish between these patterns, moreover several patterns may appear in a tissue where the predominant pattern has the clinical relevance. Studies showed that these patterns are clinically significant and are associated with patient survival [9]. Differentiating between different types of lung cancer helps for patient management and planning treatment.

1.3 Digital Pathology

Visual examination of tissue under microscope is a vital element of diagnostic medicine and it is a mechanism to investigate pathogenesis and the genetic processes such as cancer. Tissue preparation and processing to view under microscope has become increasingly automated which increased the speed at which the pathology labs can generate tissue slides.

Advances in digitization of glass slides in pathology happened much later than the digital transformation witnessed in radiology [11]. Digital technologies have enabled the digitization of these slides, allowing them to be viewed on a computer screen rather than through a microscope. Digital images present an opportunity for the development of new algorithms to automatically analyse the tissue characteristics. and allow more precise diagnoses [12].

In pathology, digital images can be used to make initial diagnoses by automated algorithms, offer second opinions for telepathology, quality assurance (e.g. re-review and proficiency testing), archiving and sharing, web accessibility, annotations, automated image analysis education and conferencing, research, marketing and business purposes, and tracking.

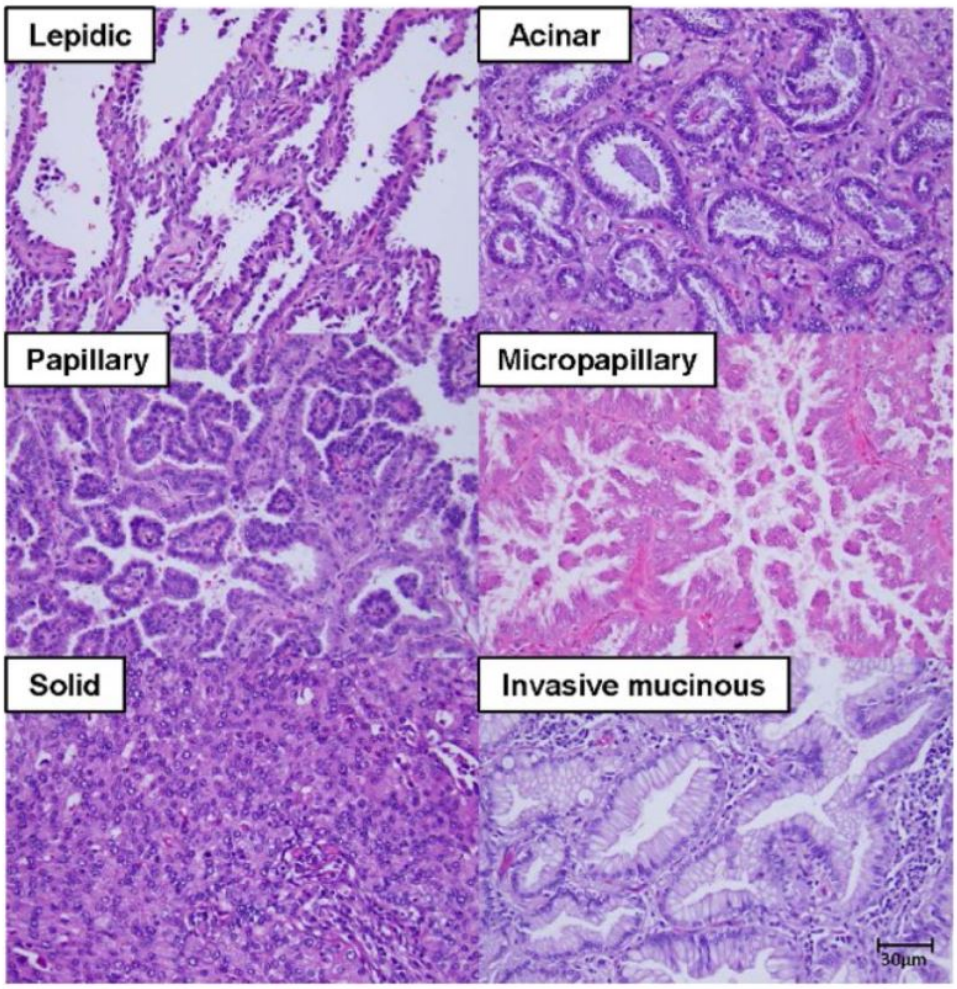

Figure 1.2: Tumour growth patterns in LUAD [10]. Lepidic growth pattern shows that tumor cells appeared to replace normal cells on alveolar walls. Acinar shows malignant glands invading a fibrous stroma, papillary adenocarcinoma consists of cuboidal tumor cells growing along fibrovascular cores in a papillary configuration. Micropapillary growth pattern shows small papillary clusters in airspace without fibrovascular cores, Solid is a sheet of nested cells, Invasive mucinous adenocarcinoma shows glandular structures filled with abundant mucin invading with an Acinar pattern. Mixtures of these patterns can appear in the tissue.

1.3.1 Whole Slide Images

Digital images at pathology slides are often referred to as whole slide images (WSIs). These images are mainly of giga-pixel size where they need roughly 50 GB memory for storing on computer. Image size of 50,000 × 50,000 is quite common. These images are typically stored in a pyramid format, where each level of the pyramid represents a different magnification level (Fig. 1.3). The highest magnification level is often 40× and the other magnification levels are presented in lower resolution with down-scaling factor of 2^n (40×, 20×, 10× etc.). Due to their large size, different compression approaches are used to compress these images such as JPEG2000 and JPEG. Even with compression, loading the whole WSI into a CPU or a GPU memory is impossible because of their limited memory capacity, therefor processing these images is challenging and mainly small images extracted from these images are used for analysis. Reading and loading these giga-pixel images requires specific libraries and software, like Openslide [13].

1.3.2 Computational Pathology Algorithms

Computation pathology is a discipline that uses computational models to analyse and process relevant data (pathology images and associated data) to assist human expert. These models aim to facilitate and improve diagnosis and patient treatment.

Computational pathology (CPath) algorithms that deal with imaging data can be categorised into the following main groups. 1) pre-processing, 2) object detection, 3) feature extraction, 4) cancer detection/diagnosis/grading, and 5) predicting patient outcome. Pre-processing algorithms in digital pathology fall into two main categories: 1) algorithms that enhance the quality of data by removing artefacts such as ink markers, tissue folding and out-of-focus regions. These artefacts can affect further analysis such as cancer detection by degrading the performance or introducing bias, therefore it is necessary to control the quality of data before any further processing. 2) Images that are captured by different scanners or at different labs can show large variations in the appearance. Optics, data acquisition algorithms and data acquisition devices can affect the color of generated image. Performance of CPath algorithms may be negatively affected by the presence of stain variation. Therefore, algorithms have been developed to standardise the stain appearance between digital images before subsequent analysis [14].

WSIs contain various objects where localising and delineating the boundaries of these objects can be helpful for performing measurements such as counting, determining the size, area, etc. Manual delineation and quantification of thousands of objects in WSIs is labour intensive and infeasible. CPath

algorithms facilitate and speed up the process of segmentation, quantification and localization of objects in WSIs, and subsequently enable the pathologists to quantitatively assess the WSIs for accurate decision. Moreover, segmentation of important objects such as nuclei and glands inside the images enable exploration of morphological features. Features that can be used for training models for cancer diagnosis or predicting patient outcome. Predicting cancer type, grade or stage is enabled through Cpath models that learn relationship between the input data and output. These models learn a mapping which can objectively predict the target of interest. They utilize the features engineered by human or they can accept raw input (via deep learning). Predicting prognosis is viable through survival models which is more complicated compared to previous tasks. The main goal in predicting the prognosis is finding features that affect the patient outcome.

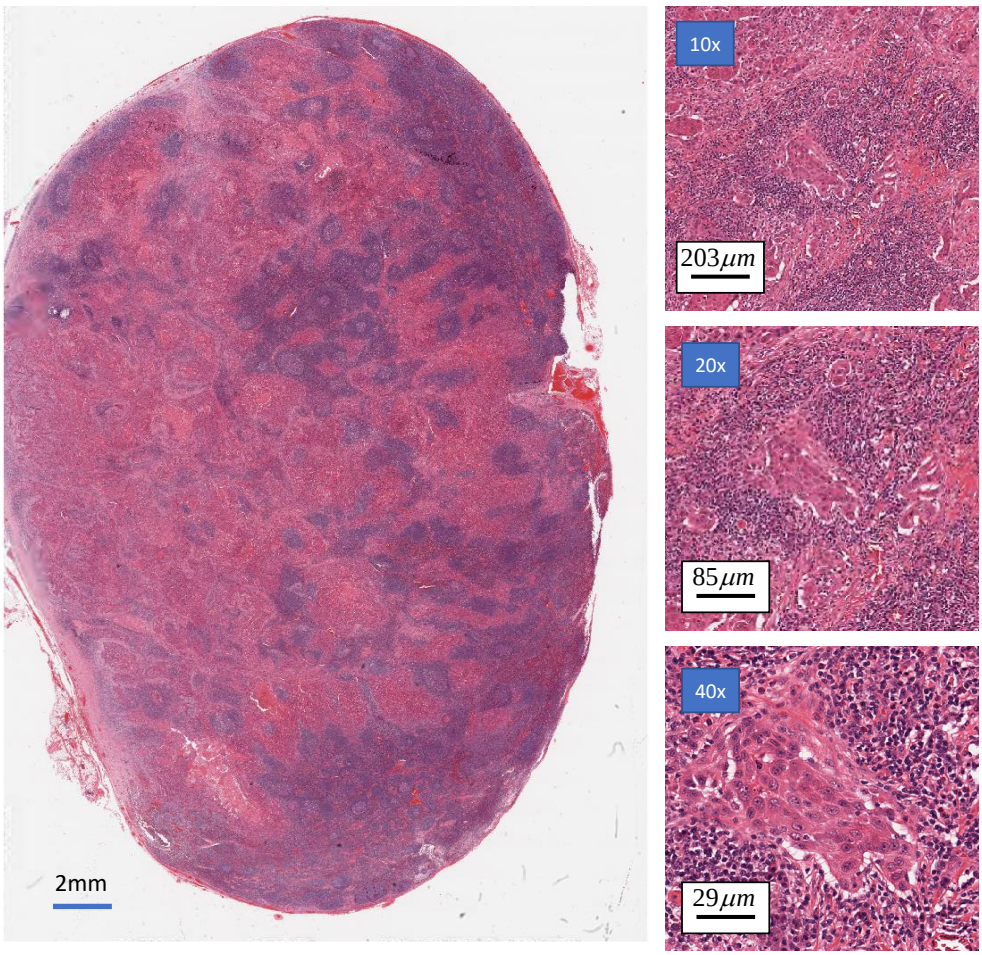

Figure 1.3: Example of whole slide image of head and neck tissue and its different magnification levels.

1.4 Deep Learning in Digital Pathology

Deep learning is a category of artificial intelligence (AI) which consists of processing layers of neural networks to learn representations of data with multiple levels of abstraction. It allows that the models learn suitable and specific features from the data and related to the task at hand. Deep learning models require very little engineering by hand in terms of both feature and model engineering. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. These methods are computationally expensive and data-hungry in nature but provided with enough data and a capable hardware, they can learn useful representations from large data sets by using the back-propagation algorithm. By back-propagation, model changes it internal parameters that are used for computing the feature/representation in each layer from the representation in the previous layer.

Owing to their powerful performance, deep networks have been extensively utilized in the pathology domain for a variety of tasks. Deep models have shown promising results for extracting clinical/biological structures in tissue. Various forms of convolutional networks have been proposed for segmenting and detecting nuclei and gland. In this thesis, we investigate this methods and we propose two methods for nuclear localization (detection and segmentation of nuclei) and gland segmentation in H&E images.

Predicting patient outcome is a topic of high interest in the area of pathology where deep models could stratify patients into the poor and good prognosis groups. In other words, deep features that are automatically extracted from networks can be utilized for survival analysis. Detecting cancer, predicting cancer sub-type, mutation, gene alteration and even detecting primary site of tumour [15] are other successful applications of deep leaning in computational pathology.

Despite its tremendous success in mapping the input to output, utilizing deep learning models have some challenges and obstacles. Three main challenges are:

- Due to their large number of parameters, they require huge amount of data to avoid over-fitting and learning the useful representations of data. And at the moment, deep learning lacks a mechanism to learn from abstract information.

- They are computationally expensive and require powerful processing hardware like GPUs. Deep models often consist of millions of parameters where powerful resources should be supplied for handling millions of

calculations required for tuning these parameters.

- Deep learning is opaque: unlike hand crafted feature that are understandable and interpretable, deep features most of the time are not easy to interpret. Deep models learn features via finding pattern and correlation from data that is often unnoticed to human. The transparency issue is important in medical domain where it is important to know how the system makes decision [16].

1.4.1 Digital Pathology Challenges

Employment of AI in digital pathology has brought impressive results. However there are some obstacles that limit the ready usage of AI for digital pathology. In the following, few of these challenges are mentioned:

Lack of Labelled Data

With the benefits that digital pathology provides, more and more hospitals and centers are going toward using WSI scanners. Despite the rapid rise in data generation, there is not enough annotation for processing of these WSIs. This is important from two aspects of model development and model evaluation. Powerful models like deep learning are highly data hungry and their performance boost is achieved when they are fed with large amount of data. To evaluate the generalizability and true performance, large annotated data sets are required which reflect the variability in the data distribution. Semi-supervised learning, domain adaptation, unsupervised learning, one/few shot learning are some approaches that to some extent may compensate for the lack of labelled data. However, still their performance can not match the fully supervised approaches. Crowd sourcing and active learning can be used for collecting annotations but the former introduce noise and the latter still needs enough annotation.

High Dimension

Digital pathology deals with large WSIs. As mentioned in Section 1.3.1, these type of images often consist of millions of pixels. Deep convolutional networks operate on much smaller images. Down-sampling the WSIs will destroy important details such as nuclei and risk high information loss. Therefore, tiling WSIs is a solution to process these large images. More precisely, each WSI is divided into smaller patches and then each patch is fed to the model separately, Fig. 1.4 shows a WSI and the patches that can be extracted. The prediction scores/probabilities of all patches are then aggregated to determine

the WSI level prediction. Losing contextual information is the main challenge in patch-wise processing of WSIs.

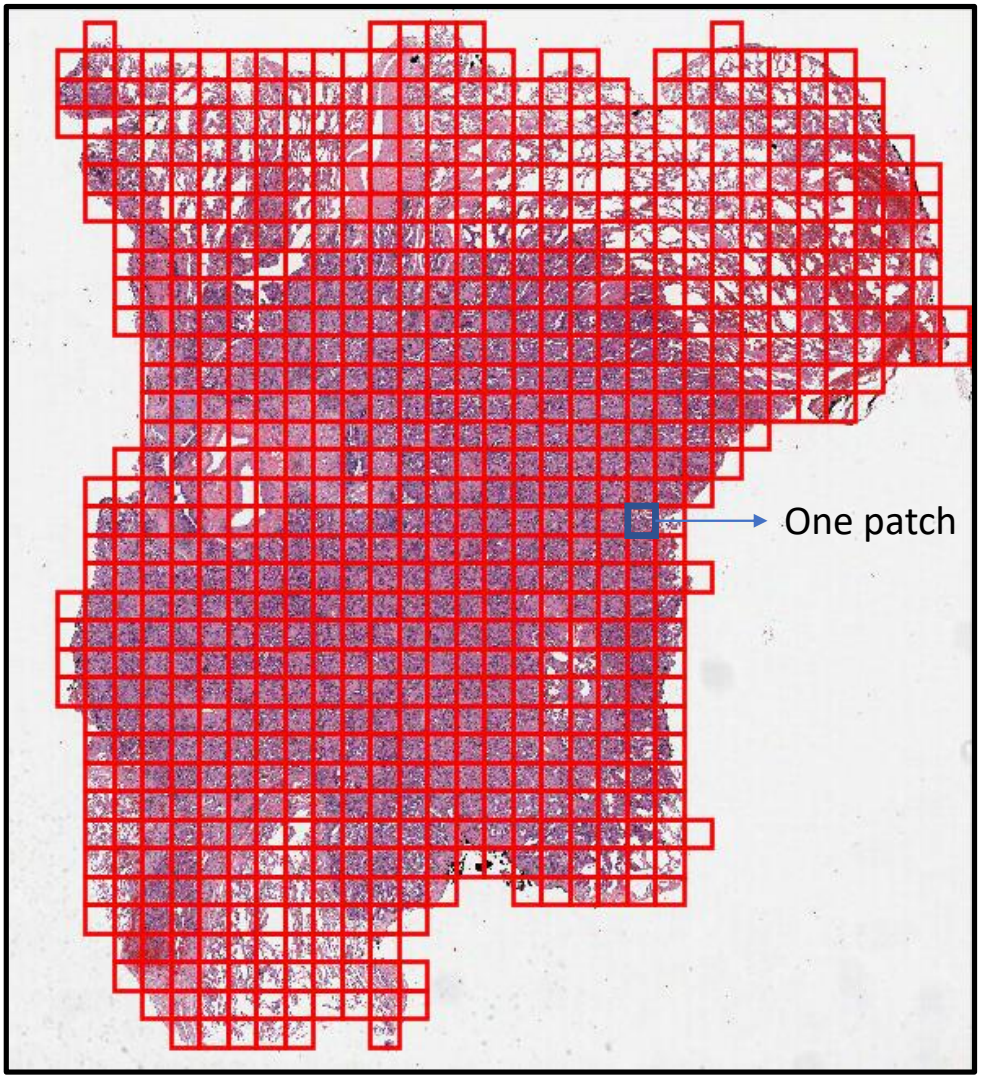

Figure 1.4: Dividing WSI into patches for subsequent processing: WSIs are inherently large images which make their processing challenging, therefore a straightforward remedy is extracting patches from them.

Computational Cost

Training deep models on the ordinary computers with Central Processing Unit (CPU) is inevitably sluggish and hence impractical. Graphical Processing Units (GPUs) through their specialized circuit enable the fast processing of pixel-based data [17]. Pathology labs are already under immense financial pressure to adopt WSI technology, and acquiring and storing gigapixel histopathological scans is a formidable challenge to the adoption of digital pathology.

Interpretability

Despite their impressive success in computer vision and scene recognition, deep models still suffer from a major drawback, which is the lack of interpretability. These models are sometimes called 'black box' due to their unknown decision making process. Researchers have proposed creative ways of explaining deep model results [18]. However, there is no established way of understanding the decisions made by them when working with histopathology images. Interpretability and transparency are important in the medical community, as experts involved in the diagnostic field usually need to justify the underlying reasons for specific decision. This is really important for deployment of deep learning algorithms in clinical practice and obtaining regulatory approvals.

Adversarial attacks

Deep neural networks can be fooled by small imperceptible variations in the image [19]. Adversarial attack is a small manipulation of input that cause the model to make a mistake[20]. Adversarial attacks are of course worrisome in medical domain, where their existence provoke us to pay more attention to noise and artefacts and more importantly to the system security. Does a noise, artifact or tissue contamination lead to an incorrect diagnosis? Can an attacker with financial incentives adversarially manipulate medical data? Designing robust AI models insensitive to adversarial attacks is challenging and there are many avenue to explore [21].

1.5 Aims and Objectives

This thesis aims to propose tools and algorithms based on deep learning to quantitatively assess pathology image tiles and WSIs. More specifically, we first propose methods for detecting and segmenting nuclei in pathology images, then we develop algorithms to facilitate collection of dense annotations from histopathology images. We also explore the effect of self supervision tasks to tackle the problem of label scarcity and finally we develop a deep model to predict survival for patients with lung cancer.

1.5.1 Main Contributions

- We propose a method based on mixture density distributions capable of detecting nuclei in images with high accuracy. We show how we can formulate mixture density networks for the task of nuclear detection.
- We develop a fully convolutional network which uses spatial and positional information for nuclear segmentation. This network utilises dense blocks

and light structure to precisely extract nuclear shapes.

- We propose NuClick, an approach for interactively segmenting nuclei and glands in histopathology images. NuClick is a generalizable approach that takes minimum human interaction and results in superior performance.
- We investigate self-supervision for learning from limited budget of annotations. Moreover, we investigate the effect of proposed domain specific self supervision tasks in the context of domain adaptation.
- We investigate the prognostic value of morphological features obtained from representative patches of WSIs. We show that multiple instance learning can be used to deal with weak labels and extracting representative patches.

1.6 Thesis Organization

Chapter 2: Nuclear Localisation in Histopathology Images

Nuclei localisation is an important task in the histology domain as it is a main step toward further analysis such as cell counting, study of cell connections, etc. This is a challenging task due to complex texture of histology image, variation in shape, and touching cells. One of the main hurdles in nuclear instance segmentation is overlapping nuclei where a smart algorithm is needed to separate each nucleus To tackle these hurdles, many approaches have been proposed in the literature where deep learning methods stand on top in terms of performance. Hence, in this chapter, we tackle the problem of nuclear localisation from two perspectives: 1) nuclear detection and 2)nuclear segmentation. For nuclear detection, we propose a novel framework for nuclei detection based on Mixture Density Networks (MDNs). These networks are suitable to map a single input to several possible outputs and we utilize this property to detect multiple seeds in a single image patch. A new modified form of a cost function is proposed for training and handling patches with missing nuclei. The probability maps of the nuclei in the individual patches are next combined to generate the final image-wide result. The experimental results show the state-of-the-art performance on complex colorectal adenocarcinoma dataset.

For nuclear segmentation, we introduce a proposal-free deep learning based framework to address these challenges. To this end, we propose a spatially-aware network (SpaNet) to capture spatial information in a multi-scale manner. A dual-head variation of the SpaNet is first utilized to predict the pixel-wise segmentation and centroid detection maps of nuclei. Based on these outputs, a single-head SpaNet predicts the positional information related to each nucleus instance. Spectral clustering method is applied on the output of the last

SpaNet, which utilizes the nuclear mask and the Gaussian-like detection map for determining the connected components and associated cluster identifiers, respectively. The output of the clustering method is the final nuclear instance segmentation mask. We applied our method on a publicly available multi-organ data set, MonuSeg, and achieved state-of-the-art performance for nuclear segmentation.

Chapter 3: A Deep Learning Framework for Interactive Segmentation of Microscopic Images

Deep learning based models generally require large amount of labeled data for precise and reliable prediction. However, collecting labeled data is expensive because it often requires expert knowledge, particularly in medical imaging domain where labels are the result of a time-consuming analysis made by one or more human experts. As nuclei, cells and glands are fundamental objects for downstream analysis in computational pathology/cytology, in this chapter we propose NuClick, a CNN-based approach to speed up collecting annotations for these objects requiring minimum interaction from the annotator. We show that for nuclei and cells in histology and cytology images, one click inside each object is enough for NuClick to yield a precise annotation. For multicellular structures such as glands, we propose a novel approach to provide the NuClick with a squiggle as a guiding signal, enabling it to segment the glandular boundaries. These supervisory signals are fed to the network as auxiliary inputs along with RGB channels. With detailed experiments, we show that NuClick is applicable to a wide range of object scales, robust against variations in the user input, adaptable to new domains, and delivers reliable annotations. As exemplar outputs of our framework, we released two datasets: 1) a dataset of lymphocyte annotations within IHC images, and 2) a dataset of segmented WBCs in blood smear images.

Chapter 4: Self-Path: Self-supervision for Classification of Pathology Images with Limited Annotations

While high-resolution pathology images lend themselves well to 'data hungry' deep learning algorithms, obtaining exhaustive annotations on these images is a major challenge. In this chapter, we propose a self-supervised CNN approach to leverage unlabeled data for learning generalizable and domain invariant representations in pathology images. The proposed approach, which we term as Self-Path, is a multi-task learning approach where the main task is tissue classification and pretext tasks are a variety of self-supervised tasks with labels inherent to the input data. We introduce novel domain specific self-supervision tasks that leverage contextual, multi-resolution and semantic

features in pathology images for semi-supervised learning and domain adaptation. We investigate the effectiveness of Self-Path on 3 different pathology datasets. Our results show that Self-Path with the domain-specific pretext tasks achieves state-of-the-art performance for semi-supervised learning when small amounts of labeled data are available. Further, we show that Self-Path improves domain adaptation for classification of histology image patches when there is no labeled data available for the target domain. This approach can potentially be employed for other applications in computational pathology, where annotation budget is often limited or large amount of unlabeled image data is available.

Chapter 5: Representative Patch Morphology Features for Predicting Lung Cancer Survival

In this chapter, we perform morphological analysis on the most discriminant patches obtained from WSIs. The patches are selected using attention block of attention-based multiple instance learning method. We apply segmentation method proposed in Chapter 3 to segment nuclei in the patches, and then we extract 68 morphological features. These features are used in the Cox model to predict the risk score. We show that this risk score is prognostically important for predicting patient outcome.

Chapter 6: Conclusions

A summary of the thesis with some potential future directions for each of the proposed methods are presented in this chapter.

Chapter 2

Nuclear Localization in Histopathology Images

2.1 Introduction

Currently many downstream analysis in digital pathology are dependant on the location of nuclei and their shapes in the pathology images. For example, their positions can be served as nodes of graph and their distance as edges for graph analysis of tumour landscape. Overall, precise localizing the nucleus in histology images is a main step for successive pathological assessment including the determination of various biomarkers, determining progesterone and estrogen receptor status (Ki-67 index), quantification of tumor immune infiltrates, counting and global/local morphological analysis [22–25]. Unfortunately, robust cell localisation is a challenging task due to nucleus clutters, large variation in shape and texture, nuclear pleomorphism, touching cells and poor image quality [26]. In addition, microscopy images often have very high resolution, which further pose a challenge on the computational resources. The evaluation of any irregularities in the appearance, morphology, and spatial organization of cell nuclei is one of the key aspects of cancer diagnosis and grading. Whole-slide histology images typically contain several hundreds of thousands of nuclei where manual detection of nucleus for further diagnostic assessment imposes a high workload on pathologists. These assessments are usually performed by visual estimation, which is labour-intensive and time-consuming, and may lead to high inter-and intra-observer variability. The ongoing of digitization in pathology has greatly contributed to the fields of pathology research and clinical practise, therefore computer assisted method can assist for automation of many assessment by processing digital images, which increases the reliability of quantitative assessments. In computer vision, object detection is defined as fitting a tight bounding box around object. Fast-RCNN [27] and YOLO [28] are two successful approaches for this task. These models are trained

based on the bounding box information (top left coordinates, width and height of bounding box), and they are vastly applied on the natural images. Cell nuclei localisation on histopathology slides requires identification of millions of densely packed small objects per image. This is in contrast to these earlier deep learning works in which usually a few dominant objects are annotated. Due to the large numbers of objects detected per image, the performance of region proposal based detectors is sub-optimal on cell detection in histology images [29]. Further, obtaining annotation of thousands of nuclei bounding boxes is impractical due to the common case of weak nuclei boundaries and high workload of pathologists. To this end, these problems are usually formulated as predicting the (x;y) coordinates of the objects' center supervised by point labels. In next section we go through some of the methods that are specifaclly designed for nuclear detection and segmentation.

Segmentation of nuclei carries more information about nucleus, where having segmentation, localization of nucleus can be obtained by extracting the centroids of segmented nucleus. However, in many cases obtaining dense prediction for training models or model evaluation is not as easy as just providing position of nucleus. So there should be models to address detection problem when segmentation dataset is not available or it is hard to obtain. In this chapter we first go through a detection model and then we elaborate on a segmentation approach.

2.2 Related Work

2.2.1 Detection

Parvin et al. [30] introduced the iterative voting methods which use oriented kernels to localize cell centers, where the voting direction and areas were updated in each iteration. Radial voting-based methods are presented for automatic cell detection on histopathology images. Qi et al [31] utilize a single path voting mechanism that is followed by clustering step. Similarly, Hafiane et al [32] detect the nuclei by clustering the segmented centers using an iterative voting algorithm. [31] utilizes single-path voting followed by mean-shift clustering and a paper by [32] that nucleus centers are detected from segmented nuclei clusters using iterative voting algorithm. Several other cell localization and segmentation methods using concave point based touching cell splitting are reported in [33] and [34], where the performances heavily rely on concave point detection.

Multiscale Laplacian-of-Gaussian (LOG) [35] and construction of concave vertex graph [33] can also be found in the literature. An approach to handle touching cells is marker-based watershed algorithm which has been used widely

in [22, 36–38]. However, due to the large variations in microscopy modality, nucleus morphology, and the inhomogeneous background, it remains to be a challenging topic for these non-learning methods. Data-driven methods utilizing hand-crafted features have also been extensively applied for cell detection due to their promising performance. Interested readers are referred to [39] for more details about methods which rely on hand crafted features and classic supervised methods.

Deep learning has shown an outstanding performance in computer vision analysis of both natural and biomedical images. Deep learning methods extract the appropriate features from an image without the need for laborious feature engineering and parameter tunning. Ciresan et al. [40] applied a deep neural network (DNN) as a pixel classifier to differentiate between mitotic and non-mitotic nuclei in breast cancer histopathology images. Xie et al. [41] proposed a structured regression convolution neural network (CNN) for nuclei detection wherein the gaussian distribution is fitted on the nucleus center to construct the probability map which is considered as an image mask, then a weighted mean squared loss is minimized via pixel-wise back-propagation. Xu et al. [42] proposed a stacked sparse autoencoder strategy to learn high level features from patches of breast histopathology images and then classify these patches as nuclear or non-nuclear. Su et al. [43] present a cell detection and segmentation algorithm using the sparse reconstruction with trivial templates and a stacked denoising autoencoder (sDAE) trained with structured labels and discriminative losses. Sirinukunwattana et al. [44] proposed a locality sensitive deep learning approach for nuclei detection in the H&E stained colorectal adenocarcinoma histology images. In this approach, a spatially constrained CNN is first employed to generate a probability map for a given input image using local information. Then the centroids of nuclei are detected by identifying local maximum intensities.

2.2.2 Segmentation

Previous methods are mainly based on region-proposal networks, like Mask-RCNN [45] and PA-Net [46], or encoder-decoder neural structures particularly U-Net model [47]. Since U-Net was not well established for separating close object in complex histology images, various methods have been introduced in the literature which concentrates on the following 4 aspects: i) modifying the network architecture to extract richer information (like CIA-Net[48]), ii) introducing auxiliary outputs to the network, the auxiliary output can be the nucleus contour or bounding box (like DCAN [49], BES-Net [50]), iii) some methods proposed CNNs that predicts distance map (or other geometrical mappings) of nuclei instances (like DR-Net [51]), and iv) taking into account

different combinations of above-mentioned variations to make their deep learning platform more robust for detecting individual objects [52]. Despite these advancements, these models lack spatial awareness which can improve instance-wise segmentation of clustered nuclei, especially in advanced stages of the tumor.

In the following, first, we introduce our approach for nucleus detection based on Mixture Density Networks (MDN) introduced by Bishop [53] for solving inverse problems, where we have multiple targets for an individual input. MDN learns the distribution of nucleus within an image hypothesizing that each nuclei has a Gaussian distribution with a maximum value on its center. Here we formalize the concept of MDN for cell detection problem. Due to MDN's flexibility to localize nucleus, we show that it has a better performance when compared with the other cell detection algorithms on a challenging colon cancer dataset.

Next we discuss our approach for nucleus instance segmentation (Spa-Net) in details. We show that not only can we achieve competitive results for nuclear instance segmentation, but also our approach is more efficient in terms of number of parameters.

Our contributions are as follows:

- We define the problem of nuclei detection as mapping a single input image patch into the probability density function (pdf) of the nuclei center, from which the observed locations have been sampled. The pdf is modeled as a Gaussian Mixture Model (GMM) and its parameters are learned via a back-propagation. In addition, a Bernoulli distribution is trained whose parameter predicts if the local patch contains any nucleus and thus the fit of the GMM is liable.

- We modified mixture density loss function in order to be well adjusted to the problem, to this end, instead of taking one target variable for each input variable, it can take multiple target variables by separating the summation term in loss function for each image, and also a binary posterior is applied on loss function to remove the images with no nucleus.

- we show that the proposed method is able to process the input image with sparsely annotated data whereas the previous methods do not consider those regions or they result in poor performance with weak annotations.

- we demonstrate the capability of algorithm to learn the distribution of nuclei center from the training data without the need to define fixed variance size for all nucleus as some methods do [41, 44].

- We propose a deep learning based proposal-free framework for nuclei

instance segmentation having low computational cost and simple post-processing.

- We propose a spatially-aware network architecture, which is equipped with a novel multi-scale dense convolutional unit.
- We propose to incorporate a nuclei detection map for estimating the number of clusters per nuclei clump.
- Our method achieves state-of-the-art results on a well-known publicly available multi-organ data set.

2.3 Nuclear Detection using Mixture Density Networks

2.3.1 Mixture Density Networks

For a general task of supervised learning our goal is to model a conditional distribution $p(t|x)$ (for image patch x and nucleus center t), which is considered Gaussian for many problems and a least square energy function is often obtained using maximum likelihood. These assumptions can lead to a poor performance in many application having plausible non-Gaussian distributions. One of such applications is one to many mapping where one input corresponds to several outputs. The assumption of having a Gaussian distribution forces the model to predict only one output discarding other target values at best. Moreover, The network prediction is the average of all target values which is incorrect [53].

Mixture Density Networks (MDN) simply combines mixture models with neural networks. In other words, each input is mapped to a mixture model distribution where the neural network estimate the parameters of mixture model. To address limitations of uni modal distributions and increase the degree of model flexibility, we can consider a general framework for modeling the conditional probability distribution by modeling it as a mixture density represented as a linear combination of kernel functions:

$$p(t|x) = \sum_{k=1}^{K} \alpha_k(x)\phi_k(t|x) \tag{2.1}$$

where K is the number of components in the mixture and α_is are mixing coefficients. We assume that kernel functions $\phi(t|x)$ are isotropic Gaussian:

$$\phi_k(t|x) = \frac{1}{(2\pi)^{c/2}\sigma_k^c(x)} \exp\left\{-\frac{\|t-\mu_k(x)\|^2}{2\sigma_k^2(x)}\right\} \tag{2.2}$$

where $\mu_k(x)$ and ${\sigma_k}^2(x)$ are the mean and the variance of the kth Gaussian, respectively, and c is the dimension of target variable. The GMM parameters can be derived from the MDN as:

$$\alpha_k(x) = \frac{\exp(z_k^{\alpha}(x))}{\sum_{l=1}^{K} \exp(z_l^{\alpha}(x))} \tag{2.3}$$

$$\mu_k(x) = z_k^{\mu}(x) \tag{2.4}$$

$$\sigma_k(x) = \exp(z_k^{\sigma}(x)) \tag{2.5}$$

where $z_k^{\alpha}(x)$, $z_k^{\mu}(x)$ and $z_k^{\sigma}(x)$ are the activations of the output layer of MDN corresponding to weight, mean and variance of $k_t h$ Gaussian in the GMM, given the input x, respectively. Softmax function in Eq. (2.3) ensures that weights of GMM sum to one and are positive values. Eq. (2.5) constraints the standard deviations to be positive.

Here, the parameters of the mixture model are considered to be functions of input image patch x. This can be achieved by using a conventional neural network as a function that takes x as input. These layers are then combined with other fully connected layers to from the Mixture Density Network (MDN), (see Fig.2.1). Building the MDN increases the number of parameters from c output to $(c + 2) \times K$, where c remains to be dimension of the output and K is the number of mixtures we are using in the model.

To define the error function, the standard negative logarithm of the maximum likelihood is used. Therefore the original loss function for the network is:

$$E = -\sum_{n=1}^{N} \ln p(t_n|x_n) = -\sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \alpha_k(x_n)\phi(t_n|x_n) \right) \tag{2.6}$$

where summation over n applies to all dataset. In the next section, we modify this cost function so that it becomes more suitable to handle image patches with multiple and/or missing nuclei.

2.3.2 Extending MDN for Nuclei Detection

For nuclei detection, deep learning approaches are either provided with small patches each containing one nuclei [42, 44] or designed as pixel wise structured logistic regression [41, 54].

Here, we formulate the cell detection as the problem of mapping one to many outputs, as each input vector (image) can have multiple variables defined as the locations (coordinates) of the nucleus. In other words, each input image

is considered to have a Gaussian mixture distribution where at the nuclei centroids the probability of each Gaussian is maximum. Formally, given image set X form domain D, our goal is to find set of:

$$D = \{(x_1, y_{11}), ..., (x_1, y_{n1}). \\ ....(x_i, y_{1i}), ..., (x_i, y_{ni})\}$$

where ni shows n_{th} nucleus in image x_i.
As x has multiple output, using Eq. (3.1) as objective function, each image should be passed ni times to the network with different labels at each forward-pass. To tackle this issue and adjust the MDN for nuclei detection, we modify the Eq. (3.1) to take one input (image patch) and all of its corresponding target coordinates of the nuclei during training at just one forward pass. To this end, the GMM is calculated ni times for each image and summation of GMMs are used to construct the objective function as shown in 2.7

$$E(W) = -\sum_{i=1}^{I}\sum_{n=1}^{N_i} \ln\left\{\sum_{k=1}^{K} \alpha_k(x_i, w) N(t_{ni}|\mu_k(x_i, w), \sigma_k^2(x_i, w))\right\} \tag{2.7}$$

where I is the number of the training images, N_i is the number of nuclei within each image and t_{ni} is the coordinate of nth nucleus centroid within the image patch i.

The network always predicts fixed number of parameters for all inputs. According to the values of weight coefficients, it is predicted if the nucleus exists in the certain location. The predicted mean of each Gaussian is used to locate nucleus. However for some input data, the predicted parameters can not be used due to unavailability of target variable. In other words, there should be a flexibility in the model to estimates the existence of targets. For nucleus detection, it means that Eq. (2.7) can only be used when all input patches contain nuclei (when we have at least one target variable for each image), whereas there are many patches with no nucleus. To address this problem, we add a Bernoulli variable $e(x_i)$ to our loss function to ignore mixture parameters for patches with no nuclei, therefore final loss function is:

$$E(W) = -y_i[\sum_{i=1}^{I}\sum_{n=1}^{N_i} \ln\left\{\sum_{k=1}^{K} \alpha_k(x_i, w) N(t_{ni}|\mu_k(x_i, w), \sigma_k^2(x_i, w))\right\}] \\ - \ln[e(x_i)^{y_i}(1 - e(x_i))^{1-y_i}] \tag{2.8}$$

Where e_i is a Bernoulli variable that specifies the probability of the patch containing any nucleus. $y_i \in [0, 1]$ is the label for each patch (for empty patch, y = 0). And if $y_i = 0$, only second part of 2.8 is back-propagated to the network.

2.3.2.1 Pointset for each nuclei

If the points lies inside the nucleus, it is considered true. However, the datasets are provided with one label for each nucleus indicating an approximate centroid location of that nucleus. To this end, during training of the network, we use a dilated point set located within 6 pixels from the nucleus center. This makes the network more robust against the variations in the locations of centroid which are not precise. We sample 10 points from a Gaussian distribution with the mean on nucleus centroid.

Conventionally, the detection using small patches generally assumes that in each patch lies one cell, hence mean square loss is sufficient to estimate one point in that patches [42, 55]. In this chapter, however, we use a different strategy: we train a neural network that decides the number and locations of cells on its own using the proposed MDN. To this end, the generated , σ, and α correspond to the nuclei position, its uncertainty, and significance of the detected location (the larger the weight, the more prominent the detection).

2.3.2.2 Network Architecture

The overall architecture of MDN comprising of a backbone and an MLP (multi layer perceptron) after that, where backbone can be any off-the-shelf CNN architecture. Here, we considered various off-the-shelf CNN backbones because of their capability to deal directly with raw images, without the need of prepossessing and an explicit features extraction process. The network is trained to capture the important aspects of the input data. By optimizing the dense representation of the input data in the feature maps, the performance of the fully connected part (MDN) is improved. In Fig. 2.1, the overall architecture of backbone is depicted where the image size of 50×50 is used as input to the model.

2.3.3 Experimental Results

2.3.3.1 Experimental Details

To optimize the network weights, an Adam optimizer with learning rate of 0.001 have been used. All models have been trained for 300 epochs with batch size of 256.

2.3.3.1.1 Backbone

Resnet [56] with 18 layers is utilized. We did not use very deep Resnet architecture as its training requires huge amount of data. Two fully connected layers are added after average-pooling to construct the whole architecture of MDN (See Fig. 2.1).

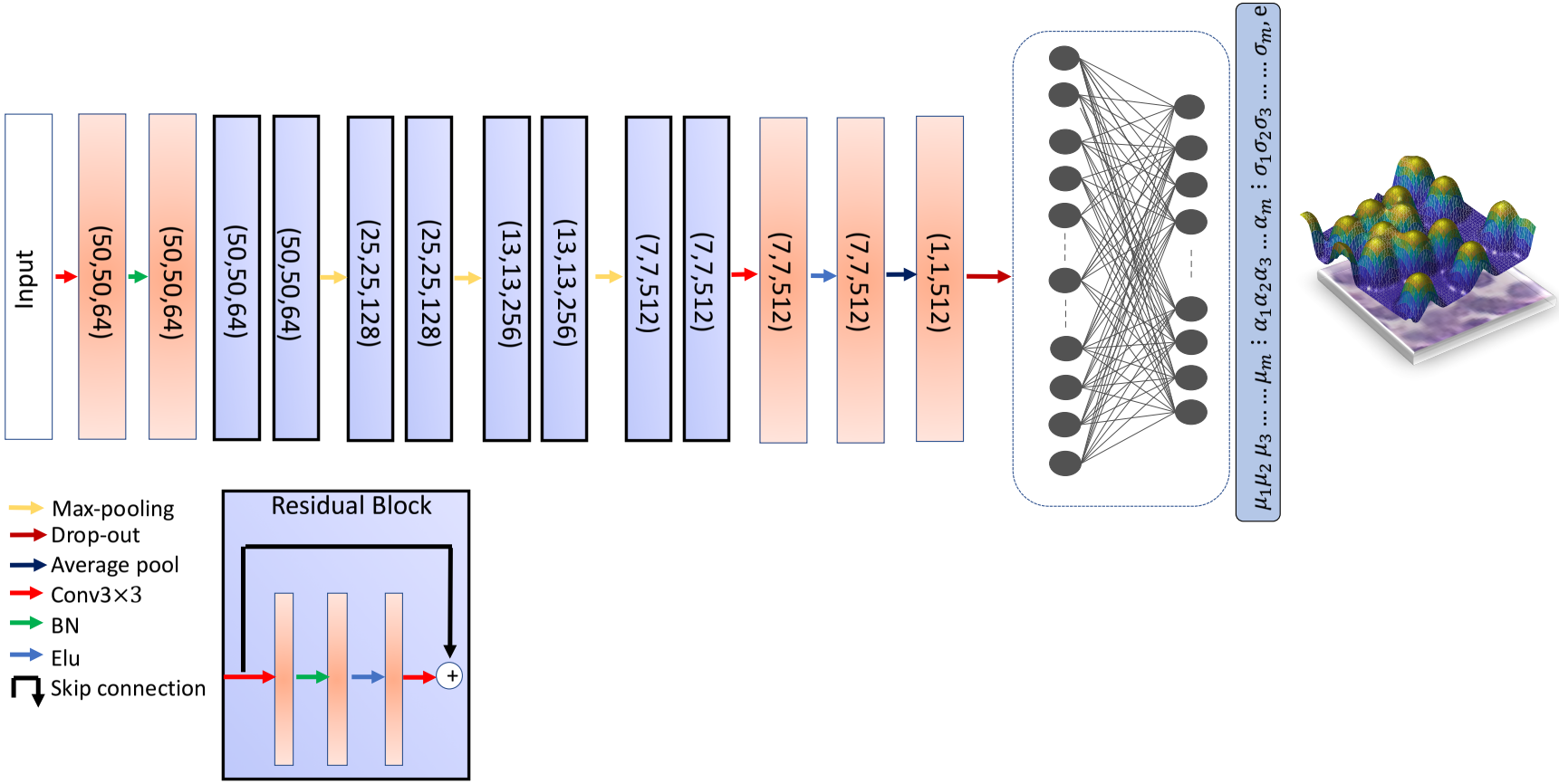

Figure 2.1: The schematic architecture of the proposed method.

To provide an appropriate input size to the network, the original images were cropped to patches of size 50×50. The network architecture consists of 2 fully connected layers (256 and $((c+2) \times K) + 1)$, respectively). We set the number of mixtures to 100, therefore the MDN should predict 401 values (for each mixture 400 values and 1 value for the Bernoulli distribution). After acquiring network predictions, the patches with no nucleus having the low value of e are ignored (threshold for e is set to 0.5). We choose the most significant Gaussians by applying a threshold of 0.001 on the mixture coefficients (α_i). Afterward, the probability maps are generated using α_i,s, σ_i,s and μ_i,s. Finally to extract the centroids of the nuclei within the remaining patches, local maxima are sought.

2.3.3.2 Dataset

For our experiments, we use the Colorectal cancer (CRC) dataset provided by [44]. It involves 100 H&E images of colorectal adenocarcinomas of size 500×500 which are cropped from CRC whole slide images. The total number of 29756 nuclei were annotated for detection purpose. The whole-slide images were obtained using an Omnyx VL120scanner All the images are obtained at 20X magnification. This dataset is randomly divided into two halves for training and testing. The cell detection on this dataset is challenging due to touching cells, blurred (or weak) cell boundaries and inhomogeneous background noise.

2.3.3.3 Metrics

Recall, precision and F1 scores are considered to evaluate the performance of nuclear detection. Similar to previous works, a circle radius of 6 pixels from each annotated nucleus is determined as the region of ground truth. True positives are the predicted locations that fall inside the ground truth circles. False positives are the predicted nuclei that are not inside the ground truth circle and false negatives are the nuclei that are not predicted by the model.

2.3.3.4 Results

In this section we investigate the performance of our proposed model on CRC dataset. For the quantitative evaluation we use the same two-fold cross validation explained in [44]. We have compared MDN with three deep models and two conventional approaches. The three deep models are SC-CNN [44]: which predicts the location and probability of each nucleus inside a small patch by using a shallow network. The probability and coordinates of nucleus are used to construct a Gaussian map for each nucleus. The ground-truth for this model is also a Gussain-like map for each nucleus. SR-CNN [41]: which is mostly similar to SC-CNN, however, they use a large patch size and

regress multiple nuclei location. More precisely, using nuclear locations, a 2D Gaussian is constructed for each nucleus where center of Gaussians are locations of nuclei and their radius (variance) is a predefined value (8 pixels). Similar to segmentation approaches, an encoder decoder model is used to predict the probability map for nuclei by using L1 loss. SSAE [42]: consists of two sparse autoencoder layers followed by a softmax classifier which is trained to distinguish between nuclear and non-nuclear patches. If classified as a nucleus, all pixels inside the output patch are assigned the value of 1, or 0 otherwise. Two conventional models are: LIPSyM [57]: that assigns high response values near the center of symmetric nuclei which is used for detection. Finally, CRImage [58] where images are thresholded and then morphological operations, distance transform and watershed is applied.

The final results are shown in Table 2.3. The algorithm has low false negatives which leads to higher recall compared to other methods. In other word, high recall highlights its performance in detecting relatively more cells compared to its counterparts. Overall the F1 score is high, which shows a good detection performance in the proposed MDN based framework. Fig. 5.3 shows the probability maps and the centroid locations along with the ground truth circles overlaid on the original images. As shown, the network could learn the locations of complex nuclei such as epithelial as well as congested area where lymphocyte nuclei lie.

The broader view of the two challenging images and their corresponding probability maps are depicted in Fig. 2.3. For detecting nuclei on the large images, we first extract patches from those large images then these patches are fed to the network to predict the location and uncertainty of nuclear locations, then we construct the heat-map for each patch, finally heat-maps are stitched together to construct the final heat map for large images. We use local maxima to localize each nucleus. For visual assessment, the annotated centroids (yellow circles) and predicted locations (red dots) are also shown in Fig. 2.3.

Due to its probabilistic output, one advantage of the proposed method is its ability to handle images with weak and sparse annotations. We demonstrate this through the following procedure. Firstly we equally divide the dataset into training and validation sets and then remove 30% of the available annotations from the training set and compare the results with SR-CNN. The quantitative results in Table 2.2, obtained using this sparsely annotated data, show that the proposed method can achieve a better performance. Our model improved the F1 score by 3% and 2% compared to SC-CNN and SR-CNN, respectively.

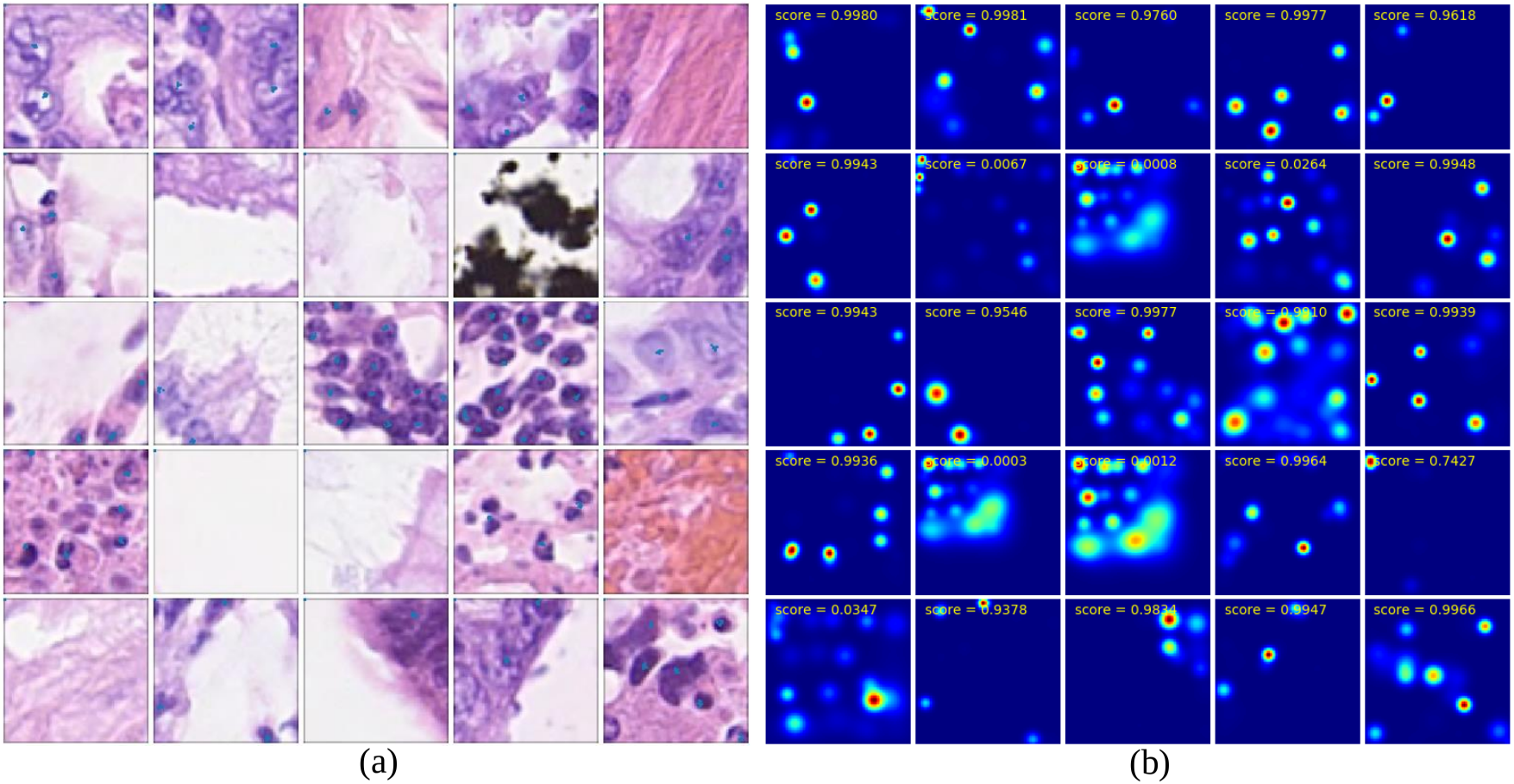

Figure 2.2: The image patches with their corresponding generated probability maps. a) The ground truth nuclei locations is overlaid on the images. b)The corresponding probability map generated using our proposed MDN. The score on top of each image is showing the probability of that patch containing any nucleus.

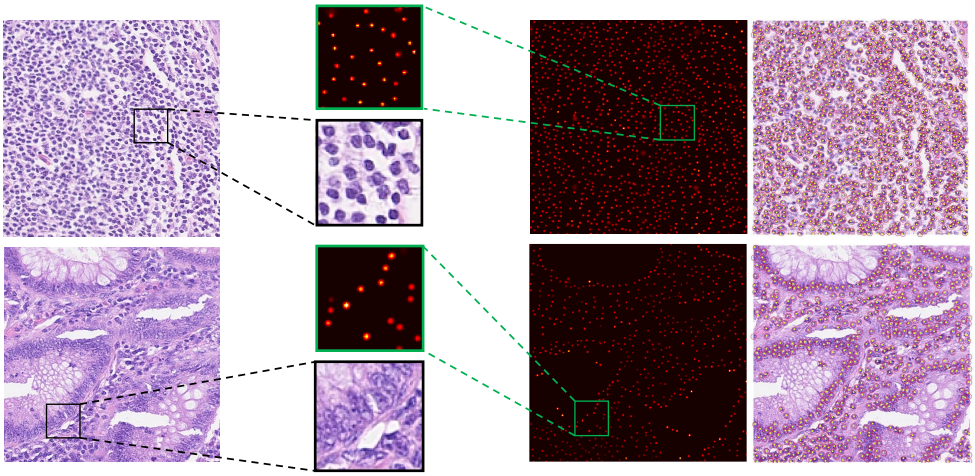

Figure 2.3: The original images on the left most column and their corresponding MDN outputs. For better visualization of congested lymphocyte nuclei (first row) and complex tumor epithelial, regions of interest are enlarged in the green boxes. The right most column shows the ground truth specified in yellow circles with detected nuclei as red dot.

Table 2.1: Comparison of precision, recall and F1 scores with other approaches.

Method	Precision	Recall	F1 score
Proposed	0.788	0.882	0.832
SC-CNN [44]	0.781	0.823	0.802
SR-CNN [41]	0.790	0.834	0.811
SSAE [42]	0.617	0.644	0.630
LIPSyM [57]	0.725	0.517	0.604
CRImage [58]	0.657	0.461	0.542

Table 2.2: Comparison of precision, recall and F1 scores using weakly annotated data.

Method	Precision	Recall	F1 score
Proposed	0.67	0.75	0.71
SR-CNN [41]	0.59	0.63	0.60

2.3.4 Discussion

In this section, we evaluate the MDN performance under variations of parameters.

2.3.4.1 Effect of Backbone

We experimented with various networks to see their effectiveness in localizing nuclei. We have considered ResNet18, Resnet50, Resnet101, DenseNet121 [59], Densenet169, VGG16 and VGG19 [60]. For fair comparison, the same values of hyper-parameters were considered in all models. The input size for all models is 50×50. We observed that large models with high number of parameters do

Table 2.3: Comparison of precision, recall and F1 scores with other approaches.

Backbone	Precision	Recall	F1 score
Resnet18	0.788	0.882	0.832
Resnet50	0.776	0.892	0.829
Resnet101	0.755	0.754	0.754
DenseNet121	0.624	0.828	0.711
VGG16	0.684	0.804	0.739
VGG19	0.676	0.793	0.729

not converge well and some times the model predict same μ_i,s for all nuclei. In table 2.3 we have shown the results for these backbones where Resnet18 depicts better performance compared to other models. For all these models number of densities (m) is set to 100.

2.3.4.2 Effect of number of components in the mixture

We also did another experiment to obtain the optimal value of K (number of Gaussian densities in GMM). To this end, we varied the value of K from 40 to 200 and increased the value by steps of 20. In CRC dataset some patches of 50 50 contains roughly 50 nuclei. Therefor we set the minimum value to 40 for this experiments. Our experiments showed that having low or very large value of K degrades the performance. Fig. 2.4 shows the the model performance (F1-score) with resnet18 backbone for different values of K. When values of K is low, the network does not have enough flexibility to locate different nuclei at different places. And when the value is too large the network struggle to learn the correct locations and sometime all densities converge to the same point.

2.4 Nuclear Instance Segmentation using a Proposal-Free Spatially Aware Deep Learning Framework

2.4.1 Methods

Our proposed method consists of predicting spatial information of each nucleus through a spatial aware CNN, and then clustering that information to construct instance-level segmentation. To achieve a reasonable spatial prediction and to estimate the number of clusters in nuclei clumps, we additionally incorporated a dual-head network for nuclei mask segmentation (semantic level) and detection maps. In this section, we firstly describe the network architecture, which is used throughout our framework. Afterward, details of employing the proposed CNN for instance segmentation will be discussed.

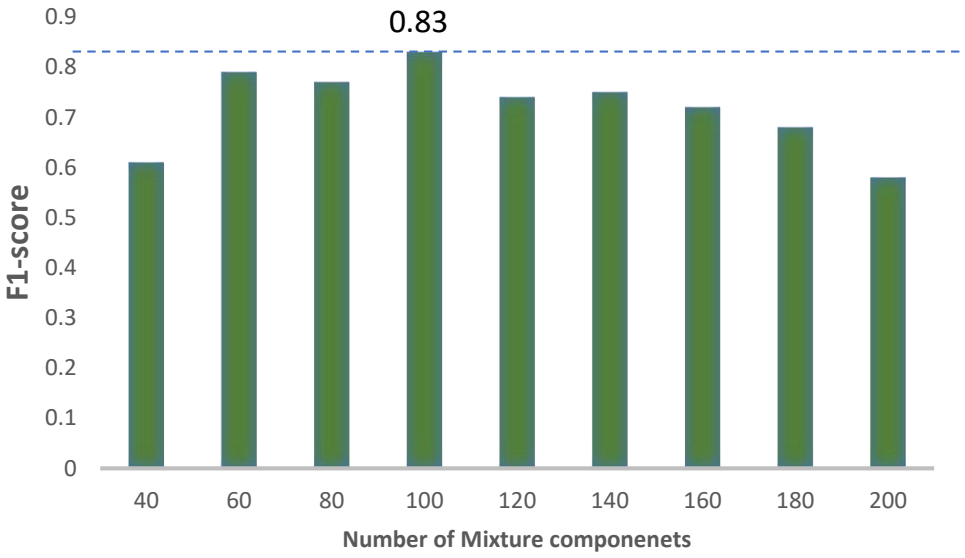

Figure 2.4: F1 score value for different number of components in the mixture: $K = 100$ could achieve best performance.

2.4.1.1 Spatially Aware Neural Network

An essential step in our proposed nuclear instance segmentation framework is predicting the positional information of each nuclei using CNNs. Conventional CNNs cannot capture positional details due to the nature of kernels. Convolutional kernels in common CNN architectures extract local features. Hence they give no intuition about the relative position of objects (detected features) in the image. To address this issue, we propose spatial information aware CNN capable of capturing positional information in all layers. By providing the network with positional information (x and y image coordinates) in the input and keeping that information available to all convolutional kernels, spatial awareness is guaranteed. Details about the positional information in the input and structuring element of the network are discussed in the following sections.

2.4.1.2 Structuring Blocks

Preserving spatial information throughout the network is feasible using our proposed multi-scale dense unit (MSDU). MSDU is a densely connected building block inspired by [59]. Unlike the ordinary dense unit, our proposed MSDU benefits from the multi-scale convolutional block (MSB) [61]. Fig. 2.5 demonstrates the configuration of a single MSB composed of four parallel convolutional blocks (convolution layer followed by batch-normalization and ReLU layers) with varying kernel size. Having the flexibility to stack convolutional blocks with varying kernel (dilation) rates allows us to obtain multi-resolution feature maps, leading to better performance. MSB Blocks are configured with

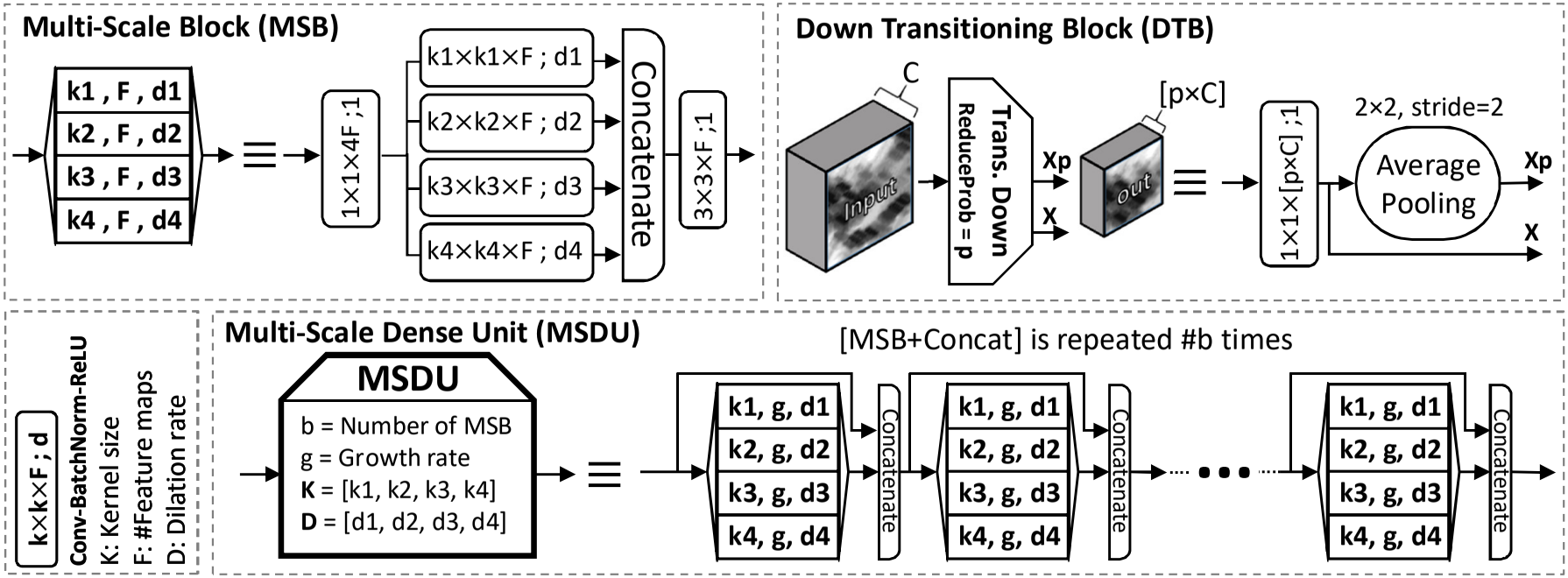

Figure 2.5: Structuring blocks used in the SpaNet architecture.

a specific number of channels (F), kernel sizes ($\mathbf{k}$), and dilation rates ($\mathbf{d}$). Each MSB block has a 1×1 and a 3×3 convolutional block in its terminals to reduce the number of processed and generated feature maps.

As depicted in Fig. 2.5, concatenation layers in the MSDU aggregate the output feature maps from their preceding MSBs. The feature aggregation property of MSDUs enables the proposed instance detection network in Fig. 2.6 to preserve the positional information (which were passed to the network's input) at all convolutional blocks throughout its path, making it a spatial aware network. An MSDU has four configuring parameters: growth rate (g), which indicates the number of feature maps generated by every MSB inside the MSDU. $\mathbf{K}$ and $\mathbf{D}$ vectors show the kernels' sizes and dilation rates of MSB blocks, and b denotes the number of MSB and concatenation pair repetitions. It has been shown that restricting the number of extracted features in each convolutional blocks (setting small growth rates) and aggregating the feature maps instead, result in better performance while reducing the computational costs [59].

Other two structuring blocks are Down Transitioning Block (DTB) and Up Transitioning Block (UTB) which down-sample and up-sample their input feature by the scale of 2, respectively. The structure of a DTB is shown in Fig. 2.5, which comprises a 1×1 convolutional block that generates $[p \times C]$ feature maps ($\mathbf{X}$). The parameter C is the number of input feature maps to the DTB, and $0 < p < 1$ is the reducing rate. DTB also consists of a 2×2 average pooling layer with a stride of 2, which will down-sample the size of feature maps in half ($\mathbf{Xp}$). UTB comprises a $2 \times 2 \times [p \times C]$ transposed convolution layer followed by batch-normalization and ReLU layers.

2.4.1.2.1 Spa-Net Architecture

The proposed spatial aware network for nuclei instance segmentation, SpaNet, is illustrated in Fig. 2.6. The main structure in SpaNet is MSDU, which is equipped with a feature aggregation property that enables positional information flows throughout the network. Feature maps in SpaNet are down-sampled three times by DTBs in the encoding path and are up-sampled accordingly by UTBs in the decoding path. Skip connections will make the feature maps in the decoding path more spatially enriched and facilitate gradient flow during training [47]. More importantly, there are some points in the network that we lose direct access to the positional information (after DTB and UTB units) where feature aggregation is not applied. As a workaround, we appropriately scaled the network input and added it in these layers via concatenation layers.

As shown in Fig. 2.6, configuring parameters of each MSDU is different, except for the growth rate (g). Other parameters are tuned based on the

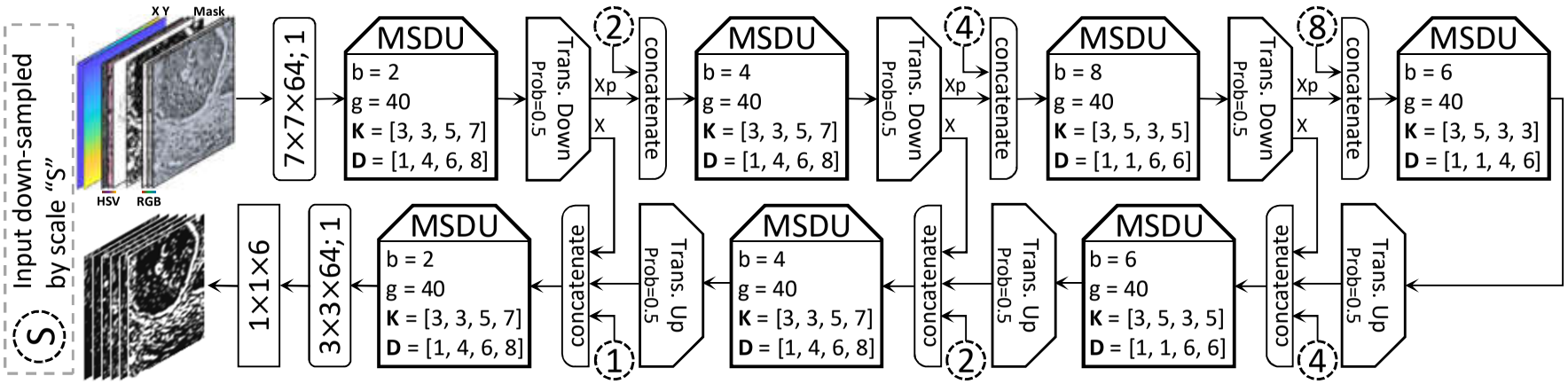

Figure 2.6: Overview of the SpaNet architecture.

MSDU position in the network. An advantage of the SpaNet is capturing small-to-large structures in all levels by appropriately setting the MSDUs' parameters. At the first level of the SpaNet where feature maps and nuclei regions are relatively large, MSDU kernel sizes and dilation rates are set to $\mathbf{K} = [3, 3, 5, 7]$ and $\mathbf{D} = [1, 4, 6, 8]$, therefore MSB convolutional kernels would have receptive field of $[3 \times 3, 9 \times 9, 25 \times 25, 49 \times 49]$ over their input feature maps. Whereas, in the final level of the encoding path where feature maps are down-sampled by a factor of 8 and are in their smallest state, MSDU kernel sizes and dilation rates are to $\mathbf{K} = [3, 5, 3, 3]$ and $\mathbf{D} = [1, 1, 4, 6]$ resulting in receptive field sizes of $[3 \times 3, 5 \times 5, 9 \times 9, 13 \times 13]$ for MSB. This means that the convolutional kernels in our proposed MSDUs can extract relevant features starting from the scale of local structures size to the scale of nucleus size. We set the parameters of MSDUs heuristically based on the nuclei diameter analysis on the available data set.

2.4.1.3 Proposal-Free Instance Segmentation

2.4.1.3.1 Segmentation and Centroid Detection

For predicting mask and position of each nucleus, a dual-head network with similar architecture to SpaNet is utilized (Fig. 2.7). One head predicts the mask of nuclei, and another head predicts the centroids. The ground truth for predicting the centroids is built by considering each nucleus as a Gaussian-Shaped function where the maximum of Gaussian occurs at the center of the nucleus. The function [62] for constructing GT for each nucleus centroid on images, $\mathbf{G}_n$, is:

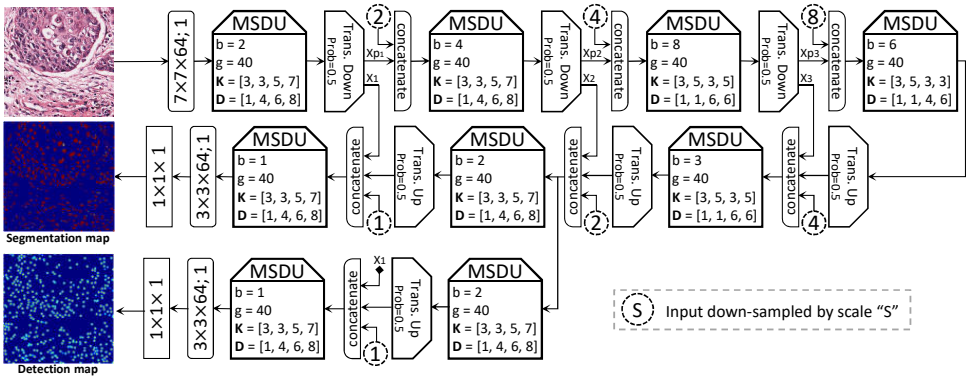

Figure 2.7: Overview of the Segmentation-Detection model based on Spa-Net architecture. For both heads, the Sigmoid activation function is utilized in the last convolutional layer.

$$\mathbf{G}_n(x, y) = \begin{cases} \frac{1}{1+\beta\|(c_{nx}, c_{ny})-(x,y)\|} & \text{if} \quad \|(c_{nx}, c_{ny}) - (x, y)\| \leq r \\ 0 & \text{elsewhere,} \end{cases} \tag{2.9}$$

where (c_{nx}, c_{ny}) and (x, y) are the coordinate of nuclei centroid and all possible coordinates of image pixels, respectively. In our experimentation β and r are, 0.01 and 8 respectively. Input to this network is RGB image, and we used smooth Jaccard and mean squared loss functions to minimize error for predicting mask and detection map, respectively. Images are first processed by this network, then the segmentation mask is attached to the RGB-HSV-XY channels to serve as the input of instance-wise network.

2.4.1.3.2 Instance Segmentation

An important part of instance segmentation is providing a ground truth (GT) that can reflect the separation between nuclei. To this end, we propose to use a GT tensor, $\mathbf{P}_{h \times w \times 6}$, that encompasses spatial information of all nuclei in the image. In $\mathbf{P}$, all pixels related to the nth nucleus, are assigned with the same feature vector of spatial information, p_n. This vector is in the form of [63]: $p_n = (c_{nx}/w, c_{ny}/h, l_{nx}/w, l_{ny}/h, r_{nx}/w, r_{ny}/h)$, where (c_{nx}, c_{ny}), (l_{nx}, l_{ny}), and (r_{nx}, r_{ny}) are the coordinates of the center, left top, and bottom right of the nth nucleus' bounding box, respectively. All the values are normalized by the width and height of bounding box, (w, h). A smoothed L_1 objective function that also ignores the background region in loss computation has been incorporated for the network optimization [63]. It is expected that the network predicts similar values for pixels belonging to the same nucleus. Note that the input to SpaNet for predicting nuclei spatial information has nine channels. The first six are made by concatenating RGB and HSV color channels, since nuclei are sometimes more distinguishable in HSV color space. The remaining 3 channels are, predicted segmentation map (achieved in the previous step), $\mathbf{M}_{\text{seg}}$, and spatial coordinate maps of pixels, $(\mathbf{M}_{\text{x}}, \mathbf{M}_{\text{y}})$. These last three channels inject the positional information to the SpaNet.

2.4.1.3.3 Post-Processing

After predicting the spatial information of nuclei instances via SpaNet, we cluster them to attain the final instance segmentation. Directly clustering the predicted maps might fail due to the large spatial domain (number of pixels) and a high number of nuclei (number of clusters) in them. Therefore, we propose to apply the clustering algorithm on nuclei clumps separately. To identify these clumps, we firstly use a threshold the segmentation maps (section 2.4.1.3.1) with a value of 0.3 and remove objects with an area smaller than 5 pixels to generate the nuclei masks. Connected components (CC) in the generated mask indicate isolated nuclei or nuclei clumps. By estimating the number of candidate nuclei (clusters) in a CC, we can start the clustering procedure. The number of clusters per CC is determined by counting the

number of local maxima in the intersection of that CC with the predicted detection map (section 2.4.1.3.1). Similar to [63] we use spectral clustering algorithm for it's effectiveness compared to other models by selecting Radial Basis Function kernel (RBF) as the affinity function. An example of network output has been shown in Fig. 2.8, the first row shows the raw prediction of segmentation and location of nuclei, and the second row shows the the prediction of 6 channels of positional information.

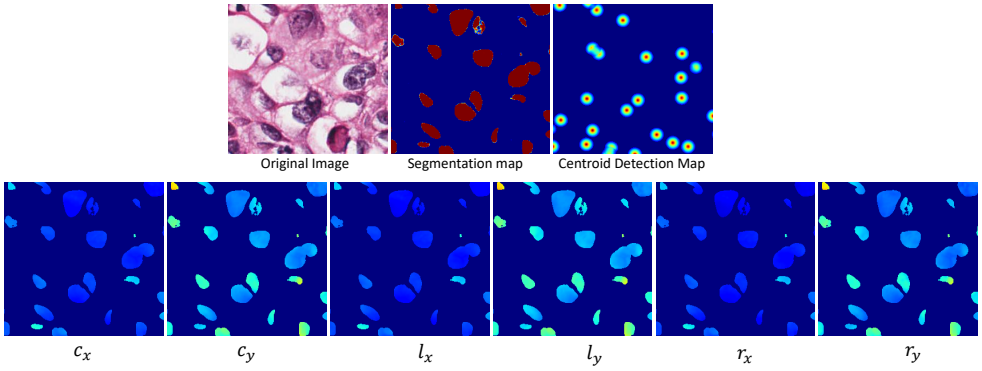

Figure 2.8: The first row shows sample outputs of the Spa-Net for Segmentation-Detection model. The second row illustrates sample outputs from the Spa-Net model which predicts positional information.

2.4.2 Results and Discussion

2.4.2.0.1 Dataset

The dataset consists of 30 H&E images (16 for training and 14 for test set) from seven different tissues. Images were obtained from The Cancer Genome Atlas (TCGA) where 1000×1000 patches were extracted from Whole Slide Images (WSIs) at 40 × magnification [64]. Only one WSI per patient was used and these images come from 18 different hospitals, which introduced another source of appearance variation due to the differences in the staining practices across labs. These seven tissues are kidney, stomach, liver, bladder, colorectal, prostate, and liver. Out of 14 test images, eight belongs to the same tissue type as the training set (seen organs) and six images are from different tissue types (unseen organs). The tissue types that are common between training and validation are: breast, liver, kidney and prostate. And the tissue types that are not provided for training are: bladder, colon and stomach. More than 21,000 nuclei are annotated in this dataset. The annotations consist of epithelial and stromal nuclei.

2.4.2.1 Evaluation metrics

For evaluating our result we have used F1-score and Aggregated Jaccard Index (AJI). F1 score is a commonly used evaluation metric which is defined as:

$$F1 = \frac{2TP}{2TP + FP + FN}$$

Where TP is the count of true positives and indicates the number of pixels that are corectly classified as the foreground. FP is the number of false positives and in the segmentation scenario it shows the number pixels background pixels that are classified as forground by the model. False negative (FN) pixels are foreground pixels that the model considered them as background (the model fails to predict those). F1-score is the same as Dice similarity. the metric does not consider the detection quality of segmentation (e.g. if objects are separated or not) and is considered for evaluating the performance of semantic segmentation. For evaluating the instance segmentation performance, detection of each object should be considered. We have used AJI for this purpose. AJI is based on Jaccard index :

$$Jacc(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

let A be the set of ground truth pixels, and B the set of predicted pixels. For calculating, Jaccard index between individual nuclei in the prediction mask and the ground truth are computed. More formally, it has two main step: 1) for each nucleus in GT (Gi), Jaccard index between that nucleus and all nuclei in the prediction mask is computed. therefore a nucleus (Si) in prediction mask that gives highest value of Jaccard can be obtained 2) we compute an aggregated intersection cardinality numerator and and aggregated union cardinality denominator for all ground truth and segmented nuclei. After associating a segmented nucleus S_i to each nucleus G_i in the ground truth, we add the contributions to the aggregated jaccard index by adding pixel count of $G_i \cap S_i$ to AJIs numerator and $G_i \cup S_i$ to the denominator. Pixel counts of all unclaimed nuclei segmented nuclei (false positives) are also added to the denomiator.

2.4.2.1.1 Networks Setup

To attain generalization and robust predictions, we followed stochastic weight averaging approach proposed in [65]. Cycling learning rate (α_i) is adopted at each iteration i as follows: $\alpha_i = (1-t_i)\alpha_1 + t_i\alpha_2$, where $t_i = (\text{mod}(i-1, c)+1)/c$, initial learning rate and final learning rate for each cycle are set to $\alpha_1 = 0.01$ and $\alpha_2 = 0.0001$, respectively, and cycling length is $c = 20$ epochs. Overall the network is trained for 100 epochs and the average of weights at the end of all

Table 2.4: Results of different methods on the nuclei instance segmentation test sets.

Method	AJI (%)		F1-score (%)	
	Seen Organ	Unseen Organ	Seen Organ	Unseen Organ
CNN3 [64]	51.54	49.89	82.26	83.22
DR [51]	55.91	56.01	-	-
DCAN[49]	60.82	54.49	82.65	82.14
PA-Net [46]	60.11	56.08	81.56	83.36
Mask-RCNN [45]	59.78	55.31	81.07	82.91
BES-Net [50]	59.06	58.23	81.18	79.52
CIA-Net [48]	61.29	63.06	82.44	**84.58**
Spa-Net (ours)	**62.39**	**63.40**	**82.81**	84.51

cycles are computed for test time prediction.

All networks in the proposed framework have been trained using the same strategy, and stochastic gradient descent has been used as an optimizer to minimize objective functions. The input patch size for all networks is 256×256. Networks for segmentation-detection and instance predictions are trained with a batch size of 2 and 4, respectively.

2.4.2.2 Augmentations

We used variety augmentation techniques to increase the model robustness against variation in appearance and shape of objects. To this end, we categorize augmentations in two parts. First category is appearance augmentation which are channel shift, contrast adjustment, applying illumination gradient and scaling intensity range. Shape augmentations are flipping horizontally and vertically, rotating image up to 40, zooming out and zooming in, shearing image, elastic deformation. These augmentations were applied randomly during training for all models.

2.4.2.2.1 Results and comparative analysis

Performance of the proposed model is compared against several deep learning based methods as reported in Table 2.4. Except the baseline method (CNN3) [64] which categories the image pixels into three classes using a CNN-based classifier, other methods in Table 2.4 (DR-Net [51], DCAN [49], BES-Net [50], and CIA-Net [48]) took a dense prediction approach and used encoder-decoder like CNN.

As deduced from the results in Table 2.4, our proposed method based on SpaNet outperforms other state-of-the-art methods. Achieving AJI of 62.39% and F1-score of 82.81% shows an improvement of 1.10% for AJI and 0.37% for F1-score metrics compared to the best performing method in the literature. The superiority of the proposed method performance can be observed in both

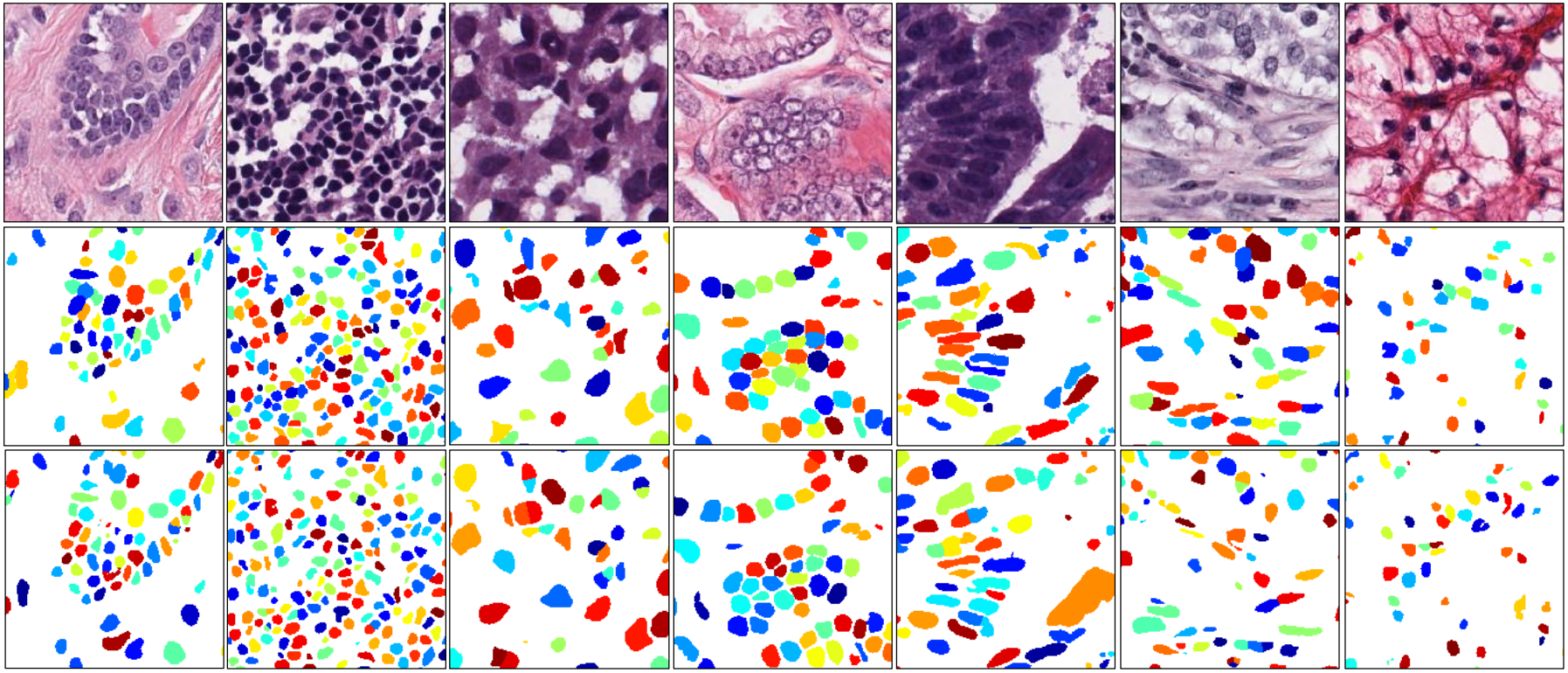

Figure 2.9: Cropped images of seven different organs with their corresponding ground truth (second row) and prediction of our proposed method (third row).

seen and unseen organs. Fig. 2.9 demonstrate the qualitative results of our method applied on all tissue types in test set.

The proposed framework offers several benefits. First, owing to the multi-scale and feature aggregation properties of MSDUs, using SpaNet architecture in this framework leads to more accurate instances' positional information. The performance of the current framework using off-the-shelf network architectures has also been shown in the Fig. 2.10 and Fig. 2.11. In Fig. 2.10, we have shown the effect of different network architectures where the network structure is replaced with SegNet [66], DeepLab [67], and U-Net [47], but all other details including labels and hyper-parameters are similar. We observe that Spa-Ne can achieve better performance on both seen and unseen test dataset when having growth rate of 43 ($g = 32$). As we increase the growth rate up to 48 the performance also improves.

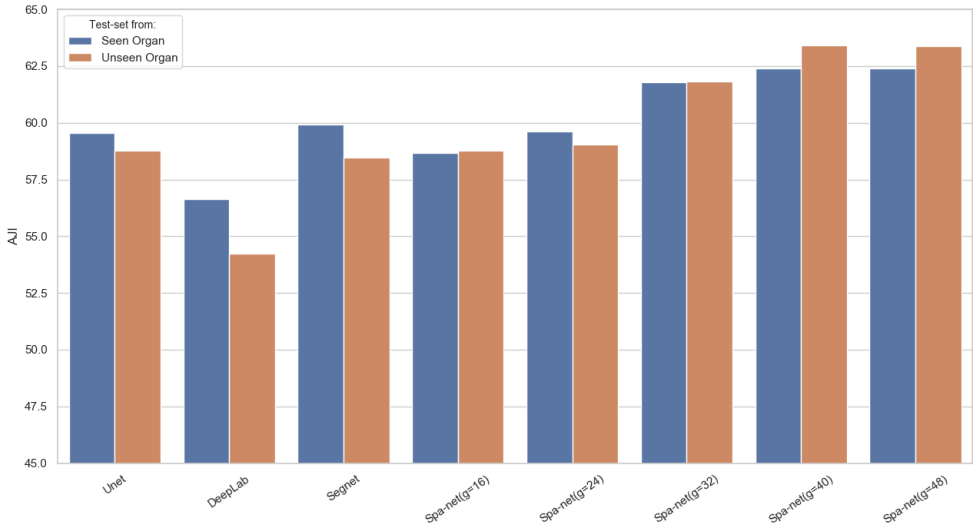

Figure 2.10: Comparison of AJI values resulted form using different networks in the proposed instance segmentation framework. These architectures are used for both segmentation-detection model and the positional information prediction network.

In Fig. 2.11, we have shown the effect of number of parameters in different models and plotted it against the Smoothed L1 loss. This plot shows that Spa-Net by increasing the number of parameters in Spa-Net the value of L1 loss is reduced. Moreover, it is inferred from this figure that, using Spa-Net with growth rate of 32 not only has fewer parameters compared to other deep structures, but also could achieve better performance (lower l1 loss). Our proposed model incorporates much less number of parameters (∼21M) in comparison with other models (∼31M for U-Net and ∼40M for CIA-Net); therefore it has a better chance to generalize on unseen data. This is an important behavior in the current application with such a small data set.

Remarkably , SpaNet models with growth rates greater than 32 outperformed other architectures in the current instance segmentation framework. However, performance of Spa-Net @40 (AJI=62.39 on seen organs and AJI=63.40 on unseen organs) and Spa-Net @48 (AJI=62.41 on seen organs and AJI=63.36 on unseen organs) are very similar. In the current research we selected the growth rate parameter equal to $g = 40$ due to the less number of network parameters and better performance on the images from unseen organs.

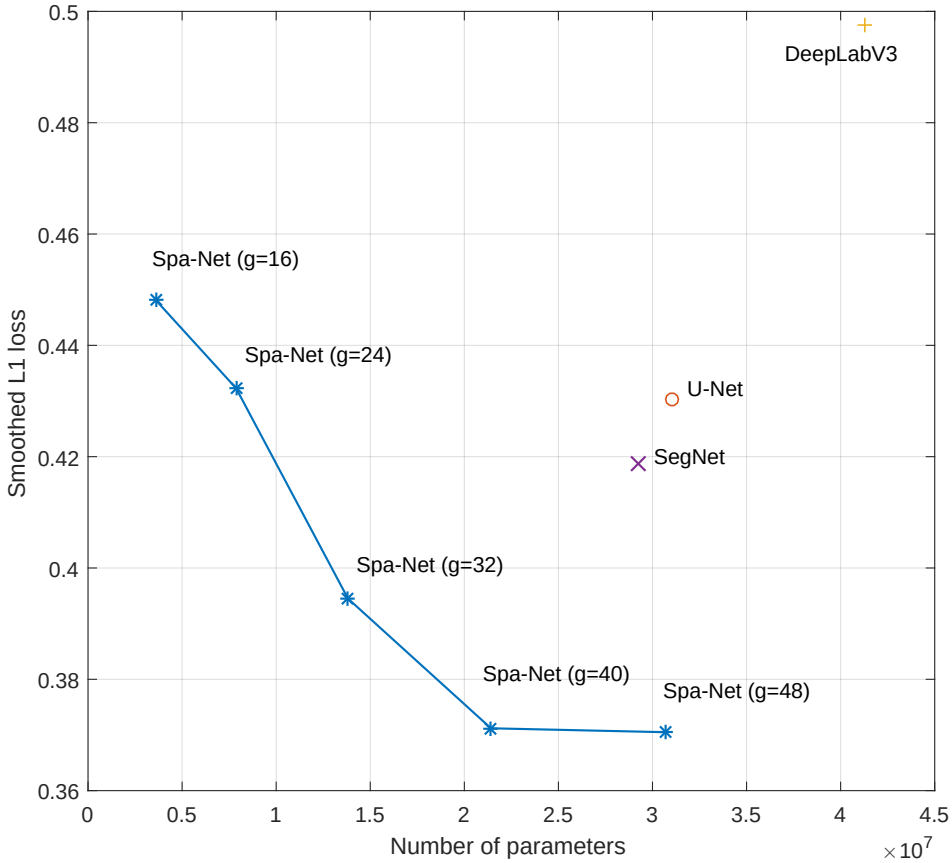

Figure 2.11: Smoothed L1 loss values for different network architectures used for positional information prediction (reported on the test set from the seen organs).

We also conducted an experiment to see the effectiveness of using a network with tree head (prediction detection, semantic segmentation and instance segmentation). To this end, we used the architecture shown in Fig. 2.12 as the triple head network. This network is compared with the model that predicts positional information and segmentation-detection in separate networks (the network that predicts segmentation-detection is shown in Fig. 2.7). We have observe (Fig. 2.7) that using two separate networks can actually give better performance.

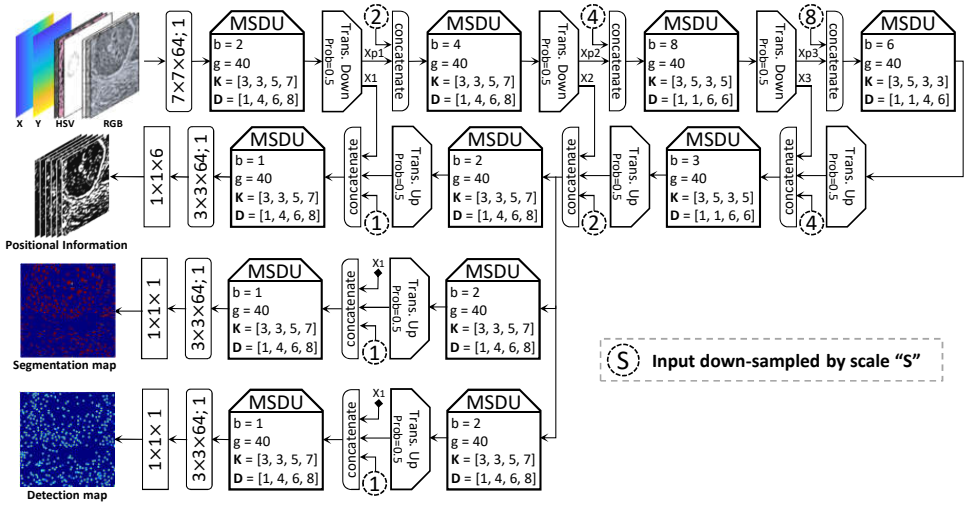

Figure 2.12: Overview of a triple head network for concurrent prediction of nuclear segmentation map, centroid detection map, and instance positional information.

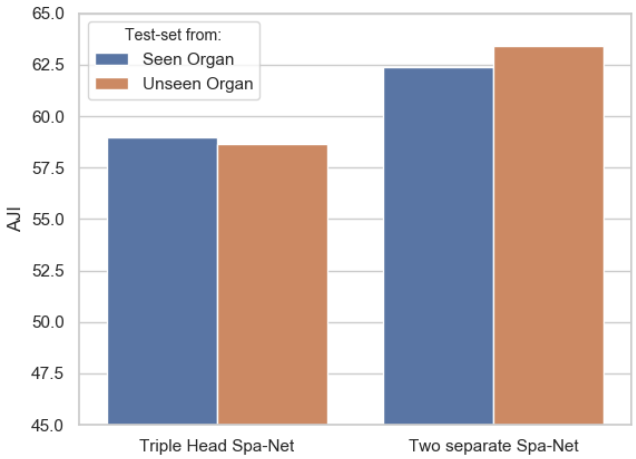

Figure 2.13: Comparison of AJI values resulted form using triple-head Spa-Net or two separate Spa-Nets (as described in the main manuscript) for prediction of segmentation map, detection map, and positional information. In both cases, growth rate is set to $g = 40$. Based on this experiment, the two separate Spa-Nets approach has been selected as the best performing model.

2.5 Summary

In this chapter, we proposed two methods for nuclear localization. First, we used a probabilistic approach for detecting nucleus. MDN has been used in literature for one to many regression tasks. Here, we proposed a framework for employing MDN for nuclei detection. The features learned using a CNN taking images as input. Then, the MDN learns the distribution of nucleus within the image patch using a mixture of Gaussian. Our method is capable of utilizing weak annotated data while preserving a good performance. Finally, we showed

that the proposed method can detect nucleus in colorectal histology images with a higher F1 score when compared to other approaches. For nuclear detection, we considered isotropic Gaussian, whereas for elongated or stromal cells non-isotropic Gaussian might be a better option. This is especially important if we want to consider our approach as a rough segmentation method.

As the second approach for nuclear localisation, we presented a proposal-free framework (Spa-Net) for nuclear instance segmentation of histology images. Prediction of segmentation map, detection map, and spatial information of nuclei were aggregated in a principled manner to obtain final instance-level segmentation. To have a precise prediction, we proposed a spatial aware network which preserves the positional information throughout the network by incorporating a novel multi-scale dense unit. We showed that Spa-Net can achieve state-of-the-art performance on a multi-organ publicly available data set.

Chapter 3

An Interactive Framework for Segmentation of Nuclei and Glands

3.1 Introduction

Automated analysis of microscopic images heavily relies on classification or segmentation of objects in the image. Starting from a robust and precise segmentation algorithm, downstream analysis subsequently will be more accurate and reliable. Deep learning (DL) approaches nowadays have state-of-the-art performance in nearly all computer vision tasks [68]. In medical images or more specifically in computational pathology (CP), DL plays an important role for tackling wide range of tasks. Despite their success, DL methods have a major problem-their data hungry nature. If they are not provided with sufficient data, they can easily over-fit on the training data, leading to poor performance on the new unseen data. In computational pathology, most models are trained on datasets that are acquired from just a small sample size of whole data distribution. These models would fail if they are applied on a new distribution (e.g new tissue types or different center that data is coming from). Hence, one needs to collect annotation from new distribution and then add it to training set to overcome false predictions.

Obtaining annotation as a target for training deep supervised models is time consuming, labour-intensive and sometimes involves expert knowledge. Particularly, for segmentation task where dense annotation is required. It is worth mentioning that in terms of performance, semi-supervised and weakly supervised methods are still far behind fully supervised methods [69]. Therefore, if one needs to build a robust and applicable segmentation algorithm, supervised methods are priority. In CP, fully automatic approaches which do not require user interactions have been extensively applied on histology images

for segmentation of different objects (e.g. cells, nuclei, glands, etc.) where DL models have shown state-of-the-art performance [70, 71].

Semi-automatic (interactive) segmentation approaches which require the user to provide an input to the system bring several advantages over fully automated approaches: 1) due to the supervisory signal as a prior to the model, interactive models lead to better performance; 2) possible mistakes can be recovered by user interactions; 3) interactive models are less sensitive to domain shift since the supervisory signal can compensate for variations in domains, in other words, interactive models are more generalizable; and 4) selective attribute of interactive models gives the flexibility to the user to choose the arbitrary instances of objects in the visual field (e.g selecting one nucleus for segmentation out of hundreds of nuclei in the ROI).

Due to generalizability power, these models can also serve as annotation tool to facilitate and speed up the annotation collection. Then these annotations can be used to train a fully automatic method for extracting the relevant feature for the task in hand. For example delineating boundaries of all nuclei, glands or any object of interest is highly labour intensive and time consuming. To be more specific, considering that annotation of one nucleus takes 10s, a visual field containing 100 nuclei takes 17 minutes to be annotated. To this end, among interactive models, approaches that require minimum user interaction are of high importance, as it not only minimizes the user effort but also speed up the process.

In this paper, by concentrating on keeping user interactions as minimum as possible , we propose a unified CNN-based framework for interactive annotation of important microscopic object in three different levels (nuclei, cells, and glands). Our model accepts minimum user interaction which is suitable for collecting annotation in histology domain.

3.2 Related Works

3.2.1 Weakly Supervised Signals for Segmentation

Numerous methods have been proposed in the literature that utilise weak labels as supervisory signals. In these methods, supervisory signal serves as an incomplete (weak) ground truth segmentation in the model output. Therefore, a desirable weakly supervised model would be a model that generalizes well on the partial supervisory signals and outputs a more complete segmentation of the desired object. These methods are not considered as interactive segmentation methods and are particularly useful when access to full image segmentation labels is limited.

For instance, [72] and [73] introduced weakly supervised nucleus segmenta-

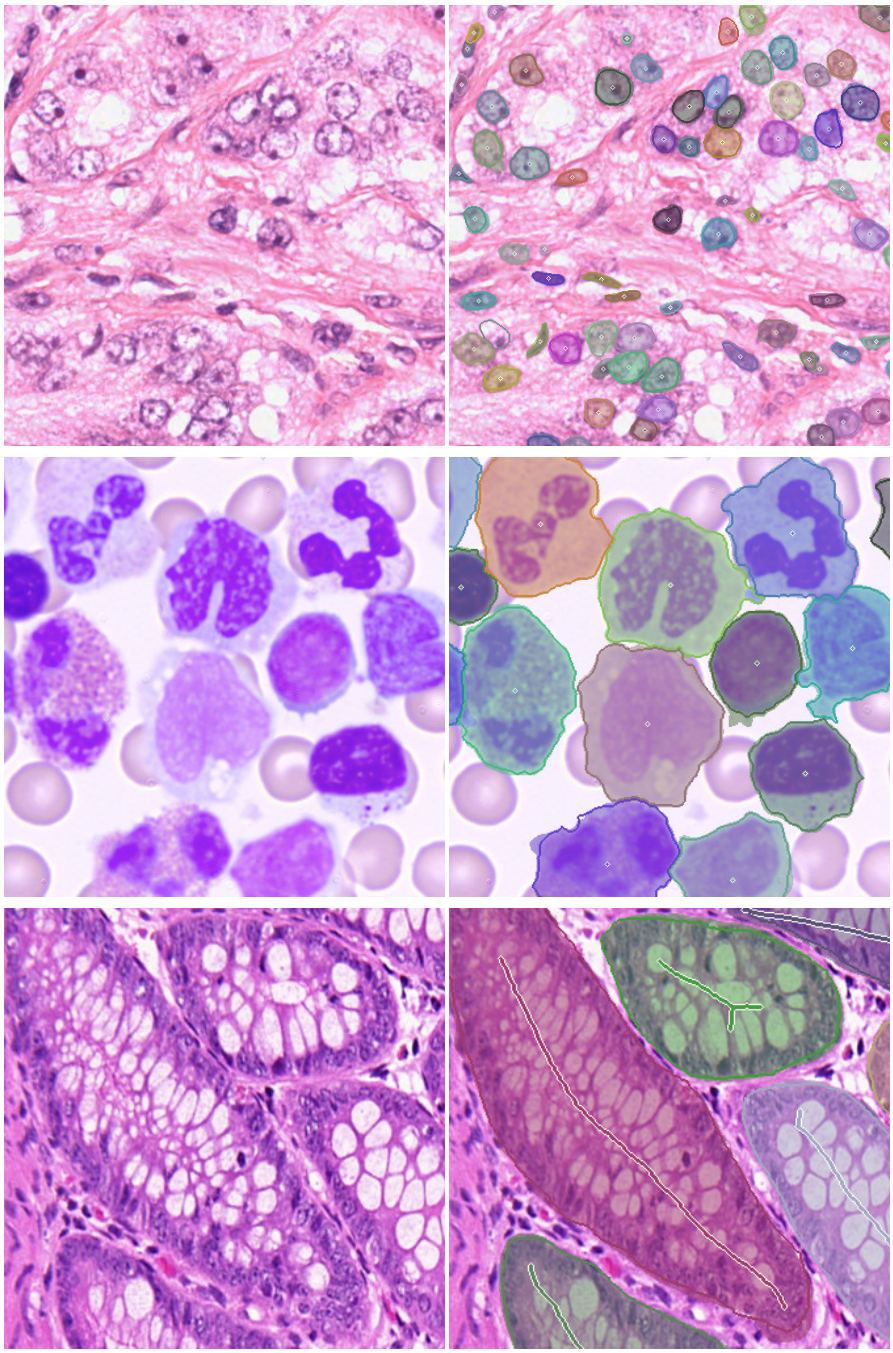

Figure 3.1: **NuClick interactive segmentation** of objects in histopathological images with different levels of complexity: nuclei (first row), cells (second row), and glands (third row). Solid stroke line around each object outlines the ground truth boundary for that object, overlaid transparent mask is the predicted segmentation region by NuClick, and points or squiggles indicate the provided guiding signal for interactive segmentation.

tion models which are trained based on nuclei centroid points instead of full segmentation masks. Several other works used image-level labels [74–77], boxes [78], noisy web labels [79, 80], point-clicks ([81–84]), and squiggles ([85, 86] as weak labels to supervise their segmentation models. Our model is analogous to methods proposed by [81] and [85] with the difference that we used points and squiggles as auxiliary guiding signals in the input of our model. Our model is fully supervised and we will show how this additional information can be used to further improve accuracy of segmentation networks on histology images.

3.2.2 Interactive segmentation

Interactive segmentation of objects has been studied for over a decade now. In many works [87–98] object segmentation is formulated as energy minimization on a graph defined over objects. In a recent unsupervised approach proposed by [99], the annotator clicks on four extreme points (left-most, right-most, top and bottom pixels), then an edge detection algorithm is applied to the whole image to extract boundaries, afterwards the shortest path between two neighboring extreme points is chosen as boundary of the object. Area within the boundaries is considered as foreground and the region outside the extreme points is considered as background for the appearance model. Grabcut [90] and Graphcut [100] are classic interactive segmentation models, which segment objects by gradually updating the appearance model. These models require the user to mark in both background and foreground regions. Although they use extensive guiding signals, they would fail if the object has blurred or complex boundaries.

In recent years, CNN models have been extensively used for interactive segmentation [99, 101–108]. A well-known example is DEXTRE [104] which utilizes extreme points as an auxiliary input to the network. First, the annotator clicks four points on the extreme positions of objects then a heat map (Gaussian map for each point where points are at the centers of Gaussians) channel is created form these clicks which is attached to the input and serves as guiding signal.

There are methods in the literature that require the user to draw a bounding box around the desired object. [97] proposed a method for interactive medical images segmentation where an object of interest is selected by drawing a bounding box around it. Then a deep network is applied on a cropped image to obtain segmentation. They also have a refinement step based on Grabcut that takes squiggles from the user to highlight the foreground and background regions. This model is applicable to single object (an organ) segmentation in CT/MRI images where this organ has similar appearance and shape in all images. However, this approach is not practical for segmentation

of multiple objects (like nuclei) or amorphous objects (like glands) in histology domain. Some methods combined bounding box annotations with Graph Convolutional Network (GCN) to achieve interactive segmentation [105–107]. In these methods the selected bounding box is cropped from the image and fed to a GCN to predict polygon/spline around object. The polygon surrounds the object then can be adjusted in an iterative manner by refining the deep model. Also, there are some hybrid methods which are based on the level sets [109]. [110] and [108] embedded the level set optimization strategy in deep network to achieve precise boundary prediction from coarse annotations.

For some objects such as nuclei, manual selection of four extreme points or drawing a bounding box is still time-consuming, considering that an image of size 512×512 can contain more that 200 nuclei. Moreover, extreme points for objects like glands are not providing sufficient guidance to delineate boundaries due to complex shape and unclear edges of such objects. In this paper, we propose to use a single click or a squiggle as the guiding signal to keep simplicity in user interactions while providing enough information. Similar to our approach is a work by [111], where the annotator needs to place two pairs of click points inside and outside of the object of interest. However, their method is limited to segmenting a single predefined object, like prostate organ in CT images unlike the multiple objects (nuclei, cell, and glands) in histology images, as is the case in this study, that mutate greatly in appearance for different cases, organs, sampling/staining methods, and diseases.

3.2.3 Interactive full image segmentation

Several methods have been proposed to interactively segment all objects within the visual field. [112] introduced Fluid Annotation, an intuitive human-machine interface for annotating the class label and delineating every object and background region in an image. An interactive version of Mask-RCNN [45] was proposed by [103] which accepts bounding box annotations and incorporates a pixel-wise loss allowing regions to compete on the common image canvas. Other older works that also segment full image are proposed by [113–116].

Our method is different from these approaches as these are designed to segment all objects in natural scenes, requiring the user to label the background region and missing instances may interfere with the segmentation of desired objects. Besides, these approaches require high degree of user interaction for each object instance (minimum of selecting 4 extreme points). However, in interactive segmentation of nuclei/cells from microscopy images, selecting four points for each object is very cumbersome. On the other hand, all above-mentioned methods are sensitive to the correct selection of extreme points which also can be very confusing for the user when he/she aims to mark a

cancerous gland in histology image with complex shape and vague boundaries. Furthermore, another problem with a full image segmentation method like [103] is that it uses Mask-RCNN backbone for RoI feature extraction which has difficulty in detecting objects with small sizes such as nuclei.

In this paper, we propose **NuClick**[1] that uses only one point for delineating nuclei and cells and a squiggle for outlining glands. For nucleus and cell segmentation, proving a dot inside nucleus and cell is fast, easy, and does not require much effort from user compared to recent methods which rely on bounding boxes around objects. For glands, drawing a squiggle inside the glands is not only much easier and user friendly for annotator but also gives more precise annotations compared to other methods. Our method is suitable for single object to full image segmentation and is applicable to a wide range of object scales, i.e. small nuclei to large glands. To avoid interference of neighboring objects in segmentation of desired object, a hybrid weighted loss function is incorporated in NuClick training.

This chapter is complementary to our previous paper [117], where we showed results of the preliminary version of NuClick and its application to nuclei, whereas here we extend its application to glands and cells. As a result of the current framework, we release two datasets of lymphocyte segmentation in Immunohistochemistry (IHC) images and segmentation mask of white blood cells (WBC) in blood sample images.

A summary of our contributions is as follows:

- We propose the first interactive deep learning framework to facilitate and speed up collecting reproducible and reliable annotation in the field of computational pathology.
- We propose a deep network model using guiding signals and multi-scale blocks for precise segmentation of microscopic objects in a range of scales.
- We propose a method based on morphological skeleton for extracting guiding signals from gland masks, capable of identifying holes in objects.
- We Incorporate a weighted hybrid loss function in the training process which helps to avoid interference of neighboring objects when segmenting the desired object.
- Performing various experiments to show the effectiveness and generalizability of the NuClick.
- We release two datasets of lymphocyte dense annotations in IHC images and touching white blood cells (WBCs) in blood sample images.

[1] Code is available at: https://github.com/navidstuv/NuClick

3.3 Methodology

3.3.1 NuClick framework overview

Unlike previous methods that use a bounding box or at least four points [89, 99, 104, 118, 119] for interactive segmentation, in our proposed interactive segmentation framework only one click inside the desired object is sufficient. We will show that our framework is easily applicable for segmenting different objects in different levels of complexity. We present a framework that is applicable for collecting segmentation for nuclei which are smallest visible objects in histology images, then cells which consist of nucleus and cytoplasm, and glands which are a group of cells. Within the current framework the minimum human interaction is utilized to segments desired object with high accuracy. The user input for nucleus and cell segmentation is as small as one click and for glands a simple squiggle would suffice.

NuClick is a supervised framework based on convolutional neural networks which uses an encoder-decoder network architecture design. In the training phase, image patches and guiding signals are fed into the network, therefore it can learn where to delineate objects when an specific guiding signal appears in the input. In the test phase, based on the user-input annotations (clicks or squiggles), image patches and guiding signal maps are generated to be fed into the network. Outputs of all patches are then gathered in a post-processing step to make the final instance segmentation map. We will explain in details all aspects of this framework in the following subsections.

3.3.2 Model architecture & loss

Efficiency of using encoder-decoder design paradigm for segmentation models has been extensively investigated in the literature and it has been shown that UNet design paradigm works the best for various medical (natural) image segmentation tasks [120, 121]. Therefore, similar to [117], an encoder-decoder architecture with multi-scale and residual blocks has been used for NuClick models, as depicted in Fig. 3.2.

As our goal is to propose a unified network architecture that segments various objects (nuclei, cells and glands), it must be capable of recognizing objects with different scales. In order to segment both small and large objects, the network must be able to capture features on various scales. Therefore, we incorporate multi-scale convolutional blocks [61] throughout the network (with specific design configurations related to the network level). Unlike other network designs e.g. DeepLab v3 [67] that only use multi-scale *atrous* convolutions in the last low-resolution layer of the encoding path, we use them in three different levels both in encoding and decoding paths. By doing this, NuClick

network is able to extract relatable semantic multi-scale features from the low-resolution feature maps and generate fine segmentation by extending the receptive fields of its convolution layers in high-resolution feature maps in the decoder part. Parameters configuration for residual and multi-scale blocks is shown on each item in the Fig. 3.2

Furthermore, using residual blocks instead of plain convolutional layers enables us to design a deeper network without risk of gradient vanishing effect ([56]). In comparison to [117], the network depth has been further increased to better deal with more complex objects like glands.

The loss function used to train NuClick is a combination of soft dice loss and weighted cross entropy. The dice loss helps to control the class imbalance and the weighted cross entropy part penalizes the loss if in the prediction map other objects rather than the desired object were present.

$$\begin{aligned}\mathcal{L} = 1 - \left(\sum_i p_i g_i + \varepsilon\right) \Big/ \left(\sum_i p_i + \sum_i g_i + \varepsilon\right) \\ -\frac{1}{n}\sum_{i=1}^{n} w_i(g_i \log p_i + (1-g_i)\log(1-p_i))\end{aligned} \tag{3.1}$$

where n is the number of pixels in the image spatial domain, p_i, g_i, and w_i are values of the prediction map, the ground-truths mask $\mathbf{G}$, and the weight map $\mathbf{W}$ at pixel i, respectively and ε is a small number. Considering that $\mathbf{G}$ has value of 1 for the desired (included) objects and 0 otherwise, its complement $\tilde{\mathbf{G}}$ has value of 1 for the undesired (excluded) objects in the image and 0 otherwise. The adaptive weight map is then defined as: $\mathbf{W} = \alpha^2\mathbf{G} + \alpha\tilde{\mathbf{G}} + 1$,where α is the adaptive factor that is defined based on areas of the included and excluded objects as follows: $\alpha = \max\left\{\sum\tilde{\mathbf{G}}\Big/\sum\mathbf{G}, 1\right\}$. This weighting scheme puts more emphasis on the object to make sure it would be completely segmented by the network while avoiding false segmentation of touching undesired objects.

3.3.3 Guiding Signals

3.3.3.1 Guiding signal for nuclei/cells

When annotator clicks inside a nucleus, a map to guide the segmentation is created, where the clicked position is set to one and the rest of pixels are set to zero which we call it *inclusion map*. In most scenarios, when more than one nucleus are clicked by the annotator (if he/she wants to have all nuclei annotated), another map is also created where positions of all nuclei except the desired nucleus/cell are set to one and the rest of pixels are set to zero, which is called *exclusion map*. When only one nucleus is clicked exclusion map is a zero map. Inclusion and exclusion maps are concatenated to RGB images to have 5 channels as the input to the network as illustrated in Fig. 3.2. The

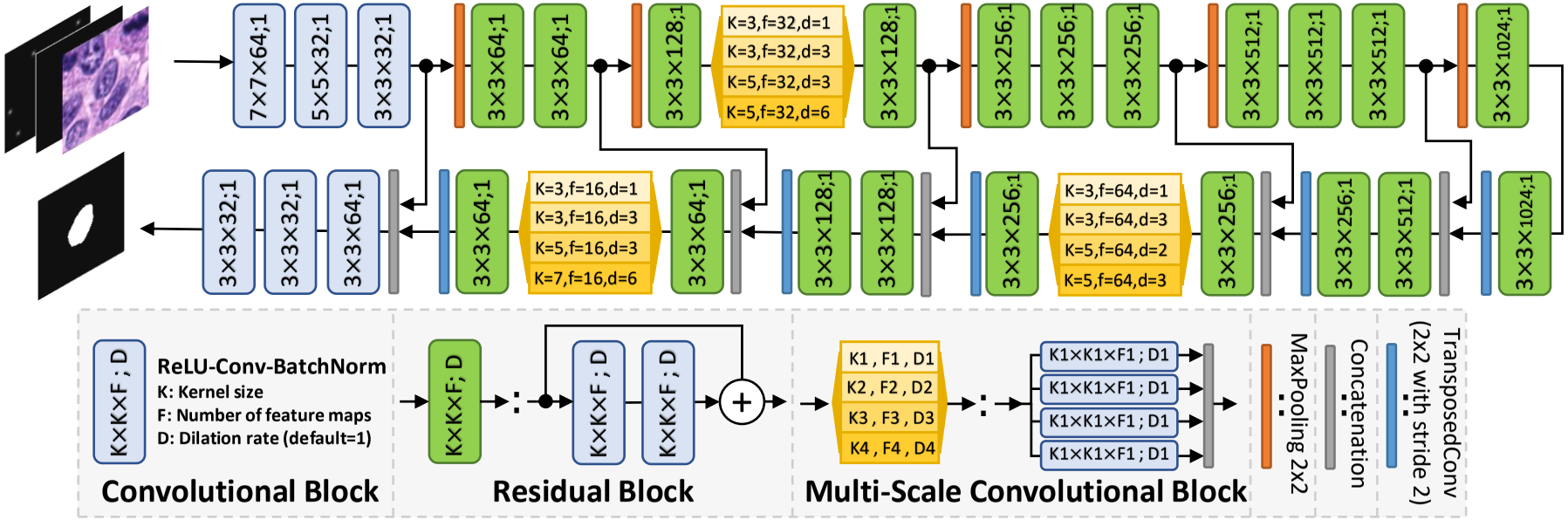

Figure 3.2: Overview of the NuClick network architecture which consists of Convolutional, Residual, and Multi-Scale convolutional blocks.

same procedure is used for creating guiding signals of cells. However, we took some considerations into the training phase of the NuClick in order to make it robust against guiding signal variations. In the following paragraphs, we will describe these techniques for both training and testing phases.

3.3.3.1.1 Training

To construct inclusion map for training, a point inside a nucleus/cell is randomly chosen. It has been taking into account that the sampled point has at least 2 pixels distance from the object boundaries. The exclusion map on the other hand is generated based on the centroid location of the rest of nuclei within the patch. Thereby, guiding signals for each patch are continuously changing during the training. Therefore the network sees variations of guiding signals in the input for each specific nuclei and will be more robust against human errors during the test. In other words the network learns to work with click points anywhere inside the desired nuclei so there is no need of clicking in the exact centroid position of the nuclei.

3.3.3.1.2 Test

At inference time, guiding signals are simply generated based on the clicked positions by the user. For each desired click point on image patch, an inclusion map and an exclusion map are generated. The exclusion map have values if user clicks on more than one nuclei/cells, otherwise it is zero. Size of information maps for nuclei and cells segmentation tasks are set to 128×128 and 256×256, respectively. For test time augmentations we can disturb the position of clicked points by 2 pixels in random direction. The importance of exclusion map is in cluttered areas where nuclei are packed together. If the user clicks on all nuclei within these areas, instances will be separated clearly. In the experimental section we will show the effect of using exclusion maps.

3.3.3.2 Guiding signal for glands

Unlike nuclei or cells, since glands are larger and more complex objects, single point does not provide strong supervisory signal to the network. Therefore, we should chose another type of guiding signal which is informative enough to guide the network and simple enough for annotator during inference. Instead of points, we propose to use squiggles. More precisely, the user provides a squiggle inside the desired gland which determines the extent and connectivity of it.

3.3.3.2.1 Training

Considering $\mathbf{M}$ as the desired ground truth (GT) mask in the output, an inclusion signal map is randomly generated as follows: First we apply a Euclidean distance transform function $D(x)$ on the mask to obtain distances of each pixel inside the mask to the closest point on the object boundaries:

$$D_{i,j}(\mathbf{M}) = \left\{ \sqrt{(\mathbf{i} - \mathbf{i_b})^2 + (\mathbf{j} - \mathbf{j_b})^2} | (\mathbf{i}, \mathbf{j}) \in \mathbf{M} \right\} \tag{3.2}$$

where i_b and j_b are the closest pixel position on the object boundary to the desired pixel position (i, j). Afterwards, we select a random threshold (τ) to apply on the distance map for generating a new mask of the object which indicates a region inside the original mask.

$$M_{i,j} = \begin{cases} 1 & if\ D_{i,j} > \tau \\ 0 & otherwise \end{cases}$$

The threshold is chosen based on the mean (μ) and standard deviation (σ) of outputs of distance function, where the interval for choosing τ is $[0, \mu + \sigma]$.

Finally, to obtain the proper guiding signal for glands, the morphological skeleton [122] of the new mask M is constructed. Note that we could have used the morphological skeleton of the original mask as the guiding signal (which does not change throughout the training phase) but that may cause the network to overfit towards learning specific shapes of skeleton and prevents it from adjusting well with annotator input. Therefore, by changing the shape of the mask, we change the guiding signal map during training. An example

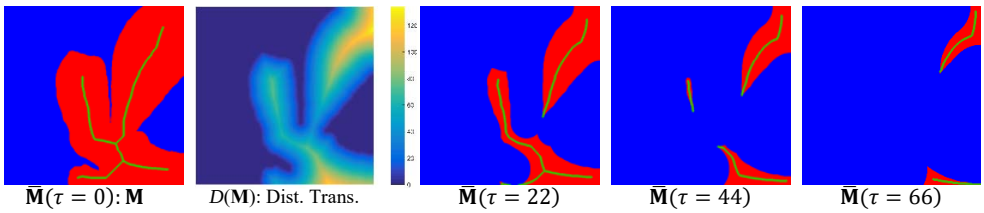

Figure 3.3: Generating supervisory signal (inclusion map) for the NuClick while training on gland dataset. The left image is the GT mask of a sample gland and $D(\mathbf{M})$ is the distance transformation of that mask. By changing the threshold value (τ), the guiding signal (skeleton of the new mask M which is specified by green color) is also changing.

of constructing map for a gland is depicted in the Fig. 3.3. In this figure, the left hand side image represents the GT of the desired gland on which its corresponding skeleton is overlaid with green color. If we use this same mask for training the network, the guiding signal would remain the exact same for all training epochs. However, based on our proposed mask changing technique,

we first calculate the distance transformation of the GT, $D(\mathbf{M})$, and then apply a threshold of τ on it to construct a new mask of M. As you can see in Fig. 3.3, by changing the the threshold value, appearance of the new mask is changing which results in different morphological skeletons as well (note the change of overlaid green colored lines with different τ values). This will make the NuClick network robust against the huge variation of guiding signals provided by the user during the test phase. The exclusion map for gland is constructed similar to nuclei/cells i.e., except one pixel from each excluding object all other pixels are set to zero.

3.3.3.2.2 Test

When running inference, the user can draw squiggles inside the glandular objects. Then patches of 512×512 are extracted from image based on the bounding box of the squiggle. If the bounding box height or width is smaller than 512, it is relaxed until height and width are 512. And if the bounding box is larger than 512 then image and corresponding squiggle maps are down-scaled to 512×512.

3.3.4 Post-processing

After marking the desired objects by the user, image patches, inclusion and exclusion maps are generated and fed into the network to predict an output segmentation for each patch. Location of each patch is stored in the first step, so it can be used later to build the final instance segmentation map.

The first step in post-processing is converting the prediction map into an initial segmentation mask by applying a threshold of 0.5. Then small objects (objects with area less than 50 pixels) are removed. Moreover, for removing extra objects except desired nucleus/cell/gland inside the mask, morphological reconstruction operator is used. To do so, the inclusion map plays the role of marker and initial segmentation is considered as the mask in morphological reconstruction.

3.4 Setups and Validation Experiments

3.4.1 Datasets

3.4.1.0.1 Gland datasets

Gland Segmentation dataset [70] (GlaS) and GRAG datasets [123, 124] are used for gland segmentation. GlaS dataset consists of 165 tiles, 85 of which for training and 80 for test. Test images of GlaS dataset are also split into to TestA and TestB. TestA was released to the participants of the GlaS challenge

one month before the submission deadline, whereas Test B was released on the final day of the challenge. Within GRAG dataset, there are a total of 213 images which is split into 173 training images and 40 test images with different cancer grades. Both of these datasets are extracted from Hematoxylin and Eosin (H&E) WSIs.

3.4.1.0.2 Nuclei dataset

MonuSeg ([71]) and CPM ([52]) datasets which contain 30 and 32 H&E images ,respectively, have been used for our experiments. 16 images of each of these datasets are used for training.

3.4.1.0.3 Cell dataset

A dataset of 2689 images consisting of touching white blood cells (WBCs) were synthetically generated for cell segmentation experiments. To this end, we used a set of 11000 manually segmented non-touching WBCs (WBC library). Selected cells are from one of the main five category of WBCs: Neutrophils, Lymphocytes, Eosinophils, Monocytes, or Basophils.

The original patches of WBCs were extracted from scans of peripheral blood samples captured by CELLNAMA LSO5 slide scanner equipped with oil immersion 100x objective lens. However, the synthesized images are designed to mimic the appearance of bone marrow samples. In other words, synthesized images should contain several (10 to 30) touching WBCs. Therefore, for generating each image a random number of cells are selected from different categories of WBC library and then they are added to a microscopic image canvas which contains only red blood cells. During the image generation each added cell is well blended into the image so its boundary looks seamless and natural. This would make the problem of touching object segmentation as hard as real images. It is worth mentioning that each WBC is augmented (deformed, resize, and rotate) before being added to the canvas. Having more than 11000 WBCs and performing cell augmentation during the image generation would guarantee that the network does not overfit on a specific WBC shape. For all datasets 20% of training images are considered as validation set.

3.4.2 Implementation Details

For our experiments, we used a work station equipped with an Intel Core i9 CPU, 128GB of RAM and two GeForce GTX 1080 Ti GPUs. All experiments were done in Keras framework with Tensorflow backend. For all applications, NuClick is trained for 200 epochs. Adam optimizer with learning rate of 3×10^{-3} and weight decay of of 5×10^{-5} was used to train the models. Batch size for nuclei, cell and gland was set to 256, 64 and 16 respectively. We used

Table 3.1: Comparison of the proposed network architecture with other models: MonuSeg dataset have been used for these experiments.

	AJI	Dice	PQ	Haus.
Unet	0.762	0.821	0.774	8.73
FCN	0.741	0.798	0.756	9.5
Segnet	0.785	0.846	0.794	8.33
NuClick W/O MS block	0.798	0.860	0.808	6.11
NuClick + 1 MS block	0.817	0.889	0.820	5.51
NuClick + 2 MS blocks	0.830	0.905	0.829	4.93
NuClick + 3 MS blocks	0.834	0.912	**0.838**	**4.05**
NuClick + 4 MS blocks	**0.835**	**0.914**	0.838	4.05

multiple augmentations as follows: random horizontal and vertical flip, brightness adjustment, contrast adjustment, sharpness adjustment, hue/saturation adjustment, color channels shuffling and adding Gaussian noise [61].

3.4.3 Metrics

For our validation study, we use metrics that have been reported in the literature for cell and gland instance segmentation. For nuclei and cells we have used AJI (Aggregated Jaccard Index) proposed by [64]: an instance based metric which calculates Jaccard index for each instance and then aggregates them, Dice coefficient: A similar metric to IoU (Intersection over Union), Hausdorff distance [70]: the distance between two polygons which is calculated per object, Detection Quality (DQ): is equivalent to $F_1 - Score$ divided by 2, SQ: is summing up IoUs for all true positive values over number of true positives and PQ: DQ×SQ [125]. For AJI, Dice, the true and false values are based on the pixel value but for DQ true and false values are based on the value of IoU. The prediction is considered true positive if IoU is higher than 0.5.

For gland segmentation, we use F1-score, Dice_{Obj}, and Hausdorff distance [70]. The true positives in F1-score are based on the thresholded IoU. Dice_{Obj} is average of dice values over all objects and Hausdorff distance here is the same as the one used for nuclei.

3.4.4 Network Selection

In this section, we investigate the effect of multi-scale blocks on NuClick network and compare its performance with other popular architectures. Ablating various choices of components in NuClick network architecture have been shown in Table 3.1. We tested our architecture with up to 4 multi-scale (MS) blocks and we observed that adding more that 3 MS blocks does not contribute significantly to the performance. It can be observed that our architecture outperforms three other popular methods (UNet by [47], SegNet by [66], and

Table 3.2: Performance of different interactive segmentation methods for nuclear segmentation on validation set of the MonuSeg dataset

Method	AJI	Dice	SQ	PQ	Haus.
Watershed	0.189	0.402	0.694	0.280	125
Region Growing	0.162	0.373	0.659	0.241	95
Active Contour	0.284	0.581	0.742	0.394	67
iFCN	0.806	0.878	0.798	0.782	7.6
LD	0.821	0.898	0.815	0.807	5.8
NuClick	**0.834**	**0.912**	**0.839**	**0.838**	**4.05**

FCN by [29]). When we use no MS block, our model is still better than all baseline models which shows the positive effect of using residual blocks. We opt to use 3 MS blocks in the final NuClick architecture because it is suggesting a competitive performance while having smaller network size.

3.4.5 Validation Experiments

Performance of NuClick framework for interactive segmentation of nuclei, cells, and glands are reported in Tables 3.2 to 3.4, respectively. For nuclei and cells, centroid of the GT masks were used to create inclusion and exclusion maps, whereas for gland segmentation, morphological skeleton of the GT masks were utilized. For comparison purposes, performance of other supervised and unsupervised interactive segmentation methods are included as well. In Tables 3.2 and 3.3, reported methods are Region Growing [126]: iteratively determines if the neighbouring pixels of an initial seed point should belong to the initial region or not (in this experiment, the seed point is GT mask centroid and the process for each nuclei/cell is repeated 30 iterations), Active Contour [127]: which iteratively evolves the level set of an initial region based on internal and external forces (the initial contour in this experiment is a circle with radius 3 pixels positioned at the GT mask centroid), marker controlled watershed [128] that is based on watershed algorithm in which number and segmentation output depends on initial seed points (in this experiment, unlike [128] that generates seed points automatically, we used GT mask centroids as seed points), interactive Fully Convolutional Network–iFCN [102]: a supervised DL based method that transfers user clicks into distance maps that are concatenated to RGB channels to be fed into a fully convolutional neural network (FCN), and Latent Diversity–LD [98]: which uses two CNNs to generate final segmentation. The first model takes the image and distance transform of two dots (inside and outside of object) to generate several diverse initial segmentation maps and the second model selects the best segmentation among them.

In Table 3.4, reported methods are Grabcut by [90]: which updates appearance model within the bounding box provided by the user, Deep GrabCut by

Table 3.3: Performance of different interactive segmentation methods for cell segmentation on test set of the WBC dataset

	AJI	Dice	SQ	PQ	Haus.
Watershed	0.153	0.351	0.431	0.148	86
Region Growing	0.145	0.322	0.414	0.129	71
Active Contour	0.219	0.491	0.522	0.198	50
iFCN	0.938	0.971	0.944	0.944	9.51
LD	0.943	0.978	0.949	0.949	8.33
NuClick	**0.954**	**0.983**	**0.958**	**0.958**	**7.45**

Table 3.4: Performance of different interactive segmentation methods for gland segmentation on test sets of the GLaS dataset

	TestA			TestB		
	F1	Dice_{Obj}	Haus.	F1	Dice_{Obj}	Haus.
Grabcut	0.462	0.431	290	0.447	0.412	312
Deep Gabcut	0.886	0.827	51	0.853	0.810	57
DEXTRE	0.911	0.841	43	0.904	0.829	49
Mask-RCNN	0.944	0.875	35	0.919	0.856	41
BIFseg	0.958	0.889	28	0.921	0.864	38
NuClick	**1.000**	**0.956**	**15**	**1.000**	**0.951**	**21**

[101]: which converts the bounding box provided by the user into a distance map that is concatenated to RGB image as the input of a deep learning model, DEXTRE [104]: a supervised deep learning based method which is mentioned in the Section 3.2.2 and accepts four extreme points of glands as input (extreme points are extracted based on each object GT mask), and a Mask-RCNN based approach proposed by [103]: where the bounding box is also used as the input to the Mask-RCNN. [103] also added a instance-aware loss measured at the pixel level to the Mask-RCNN loss. We also compared our method for gland segmentation with BIFseg [97] that needs user to crop the object of interest by drawing bounding box around it. The cropped region is then resized and fed into a resolution-preserving CNN to predict the output segmentation. [97] also used a refinement step which is not included in our implementation.

For GrabCut, Deep GrabCut, BIFseg, and Mask-RCNN approaches the bounding box for each object is selected based on its GT mask. For iFCN and LD methods, positive point (point inside the object) is selected according to the centroid of each nucleus and negative click is a random point outside the desired object.

Based on Table 3.2, NuClick achieved AJI score of 0.834, Dice value of 0.912, and PQ value of 0.838 which outperformed all other methods for nuclear segmentation on MonuSeg dataset. Performance gap between NuClick and other unsupervised methods is very high (for example in comparison

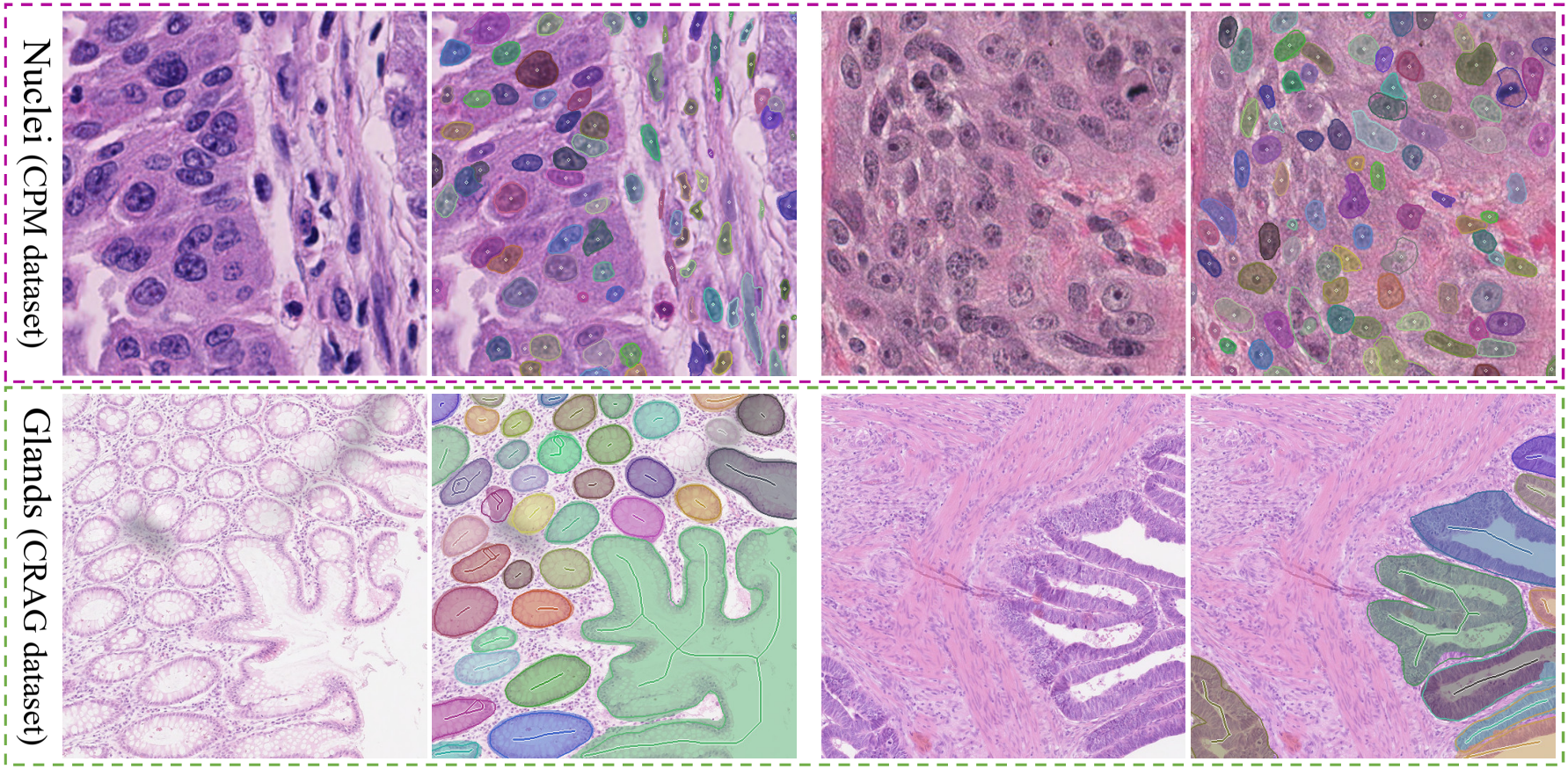

Figure 3.4: **Generalizability of the NuClick:** The first row shows results of the NuClick on the CPM dataset for nuclei segmentation (where the network was trained on the MoNuSeg dataset). The second row illustrates two samples of gland segmentation task from the CRAG dataset where the model was trained on the GLaS dataset. Solid stroke line around each object outlines the ground truth boundary for that object, overlaid transparent mask is the predicted segmentation region by the NuClick, and points or squiggles indicate the provided guiding signal for interactive segmentation. (Best viewed in color)

with Watershed method, NuClick achieves a 0.645 higher AJI). Extreme low evaluation values achieved by unsupervised metrics indicate that they are not suitable for intricate task of nuclear segmentation, even if they are fed with GT markers. There is also iFCN ([102]), a deep learning based method in Table 3.2 that is trained based on the clicked dots inside and outside of objects. However, NuClick performs better than iFCN for all AJI, Dice, and PQ metrics by margin of 2.8%, 3.4%, and 5.6%, respectively, which is a considerable boost. For the other CNN based method in Table 3.2, LD method, NuClick advantage over all metrics is also evident.

The same performance trend can be seen for both cell and gland segmentation tasks in Tables 3.3 and 3.4. For the cell segmentation task, NuClick was able to segment touching WBCs from synthesized dense blood smear images quite perfectly. Our proposed method achieves AJI, Dice, and PQ values of 0.954, 0.983, and 0.958, respectively, which indicates remarkable performance of the NuClick in cell segmentation.

Validation results of our algorithm on two test sets from GlaS dataset (testA and testB) are reported in Table 3.4 alongside the results of 4 supervised deep learning based algorithms and an unsupervised method (Grabcut). Markers used for Grabcut are the same as ones that we used for NuClick. Based on Table 3.4 our proposed method is able to outperform all other methods for gland segmentation in both testA and testB datasets by a large margine. For testB, NuClick achieves F1-score of 1.0, Dice similarity coefficient of 0.951, and Hausdorff distance of 21, which compared to the best performing supervised method (BIFseg) shows 7.9%, 8.7%, and 17 pixels improvement, respectively. The F1-score value of 1.0 achieved for NuClick framework in gland segmentation experiment expresses that all of desired objects in all images are segmented well enough. As expected, unsupervised methods, like Grabcut, perform much worse in comparison to supervised method for gland segmentation. Quantitatively, our proposed framework shows 55.3% and 53.9% improvement compared to Grabcut in terms of F1-score and Dice similarity coefficients. The reason for the advantage of NuClick over other methods mainly lies in its squiggle-based guiding signal which is able to efficiently mark the extent of big, complex, and hollow objects. It is further discussed in Section 3.5.

Methods like DEXTRE, BIFseg, and Mask-RCNN are not evaluated for interactive nucleus/cell segmentation, because they may be cumbersome to apply in this case. These methods need four click points on the boundaries of nucleus/cell (or drawing a bounding box for each of them) which is still labour-intensive as there may be a large number of nuclei/cells within an image.

Segmentation quality for three samples are depicted in Fig. 3.1. In this figure, the first, second, and third rows belong to a sample drawn from MoNuSeg, WBC, and GLaS validation sets. The left column of Fig. 3.1 shows original

images and images on the right column contains GT boundaries, segmentation mask, and guiding signals (markers) overlaid on them. Guiding signals for nuclei and cell segmentation are simple clicks inside each object (indicated by diamond-shape points on the images) while for glands (the third row) guiding signals are squiggles. In all exemplars, extent of the prediction masks (indicated by overlaid transparent colored region) are very close to the GT boundaries (indicated by solid strokes around each object).

3.5 Discussions

In order to gain better insights into the performance and capabilities of the NuClick, we designed several evaluation experiments. In this section we will discuss different evaluation experiments for NuClick. First we will assess the generalizability of the proposed framework, then we will discuss how it can adapt to new domains without further training, after that the reliability of NuClick output segmentation is studied. Moreover, sensitivity of output segmentation to variations in the guiding signals is also addressed in the following subsections.

3.5.1 Generalization study

To show the generalizability of the NuClick across an unseen datasets, we designed an experiment in which NuClick is trained on the training set of a specific dataset and then evaluated on the validation set of another dataset but within the same domain. Availability of different labeled nuclei and gland datasets allow us to better show the generalizability of our proposed framework across different dataset and different tasks.

To assess the generalizability upon nuclei segmentation, two experiments were done. In one experiment, NuClick was trained on training set of MoNuSeg dataset and then evaluated on the validation set of CPM dataset. In another experiment this process was done contrariwise where CPM training set was used for training the NuClick and MoNuSeg testing set was used for the evaluation. Evaluation results of this study are reported in the first two rows of Table 3.5. From this table we can conclude that NuClick can generalize well across datasets because it gains high values for evaluation metrics when predicting images from dataset that was not included in its training. For example, when NuClick is trained on the MoNuSeg training set, Dice and SQ evaluation metrics resulted for CPM validation set are 0.908 and 0.821, respectively, which are very close to the values reported for evaluating the MoNuSeg validation set using the same model i.e., Dice of 0.912 and SQ of 0.839 in Table 3.2. This closeness for two different datasets using the same

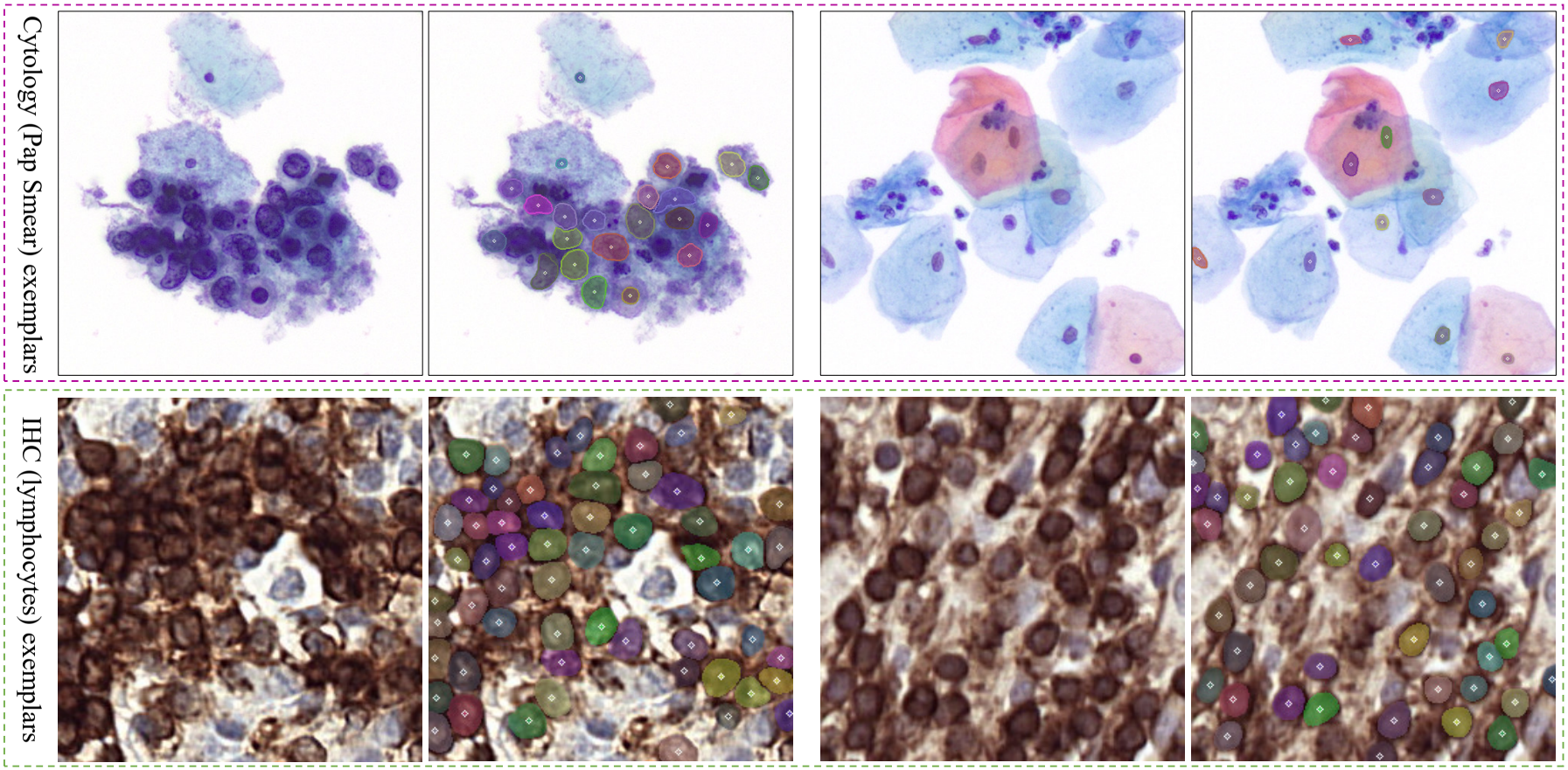

Figure 3.5: **Domain adaptability of NuClick:** nuclei from unseen domains (Pap Smear sample in the first row and IHC stained sample in the second tow) are successfully segmented using the NuClick which was trained on MoNuSeg dataset. In all images, solid stroke line around each object outlines the ground truth boundary for that object (except for IHC samples, for which ground truth masks are unavailable), overlaid transparent mask is the predicted segmentation region by NuClick, and points indicate the provided guiding signal for interactive segmentation. (Best viewed in color)

Table 3.5: Results of generalization study across different datasets for interactive nuclei and gland segmentation

	Train	Test	Dice	SQ	Dice_{Obj}	Haus.
Nuclei	MoNuSeg	CPM	0.908	0.821	-	-
	CPM	MoNuSeg	0.892	0.811	-	-
Gland	GLaS	CRAG	-	-	0.932	31
	CRAG	GLaSA	-	-	0.944	28
	CRAG	GLaSB	-	-	0.938	30

model supports our claim about generalizability of the NuClick.

Similarly, to test the generalizability of the NuClick when working on gland segmentation task, it has been trained on one gland dataset and tested on validation images from another gland dataset. As GlaS test set is divided into TestA and TestB, when NuClick is trained on CRAG, it has been test on testA and testB of GlaS (named as GlaSA and GlaSB in Table 3.5). High values of Dice_{Obj} metric and low values for Hasdroff distances also supports the generalizability of NuClick framework for gland segmentation task as well.

To provide visual evidence for this claim, we illustrated two nuclear segmentation samples from CPM validation set (resulted using a model trained on MoNuSeg dataset) and two gland segmentation samples from CRAG validation set (resulted using a model trained on GLaS dataset) in Fig. 3.4. In all cases NuClick was able to successfully segment the desired objects with high accuracy. In all images of Fig. 3.4 different overlaid colors corresponds to different object instances, solid stroke lines indicate GT boundaries, transparent color masks show the predicted segmentation region, and other point or squiggle markers representing guiding signals for interactive segmentation.

3.5.2 Domain adaptation study

To assess the performance of the NuClick on unseen samples from different data domains, we trained it on MoNuSeg dataset which contains labeled nuclei from histopathological images and then used the trained model to segment nuclei in cytology and immunohistochemistry (IHC) samples.

In the cytology case, a dataset of 42 FoVs were captured from 10 different Pap Smear samples using CELLNAMA LSO5 slide scanner and 20x objective lens. These samples contain overlapping cervical cells, inflammatory cells, mucus, blood cells and debris. Our desired objects from these images are nuclei of cervical cells. All nuclei from cervical cells in the available dataset of Pap Smear images were manually segmented with the help of a cytotechnologist. Having the GT segmentation for nuclei, we can use their centroid to apply the NuClick on them (perform pseudo-interactive segmentation) and also evaluate the results quantitatively, as reported in Table 3.6. High values of evaluation

Table 3.6: Performance of the NuClick framework on segmenting nuclei in images from an unseen domain (Pap Smear)

Method	AJI	Dice	SQ	DQ	PQ
NuClick	0.934	0.965	0.933	0.997	0.931

metrics reported in Table 3.6 shows how well NuClick can perform on images from a new unseen domain like Pap Smear samples. Some visual examples are also provided in Fig. 3.5 to support this claim. As illustrated in the first row of Fig. 3.5, NuClick was able to segment touching nuclei (in very dense cervical cell groups) from Pap Smear samples with high precision. It is able to handle nuclei with different sizes and various background appearances.

For the IHC images, we utilized NuClick to delineate lymphocytes. The dataset we have used for this section is a set of 441 patches with size of 256×256 extracted from LYON19 dataset. LYON19 is scientific challenge on lymphocyte detection from images of IHC samples. In this dataset samples are taken from breast, colon or prostate organs and are then stained with an antibody against CD3 or CD8 [129] (membrane of lymphocyte would appear brownish in the resulting staining). However, for LYON19 challenge organizers did not release any instance segmentation/detection GTs alongside the image ROIs. Therefore, we can not assess the performance of NuClick segmentation on this dataset quantitatively. However, the quality of segmentation is very desirable based on the depicted results for two random cases in the second row of Fig. 3.5. Example augmentations in Fig. 3.5 are achieved by clicks of a non-expert user inside lymphocytes (based on his imperfect assumptions). As it is shown in Fig. 3.5, NuClick is able to adequately segment touching nuclei even in extremely cluttered areas of images from an unseen domain. These resulting instance masks were actually used to train an automatic nuclei instance segmentation network, SpaNet [130], which helped us achieve the first rank in LYON19 challenge. In other words, we approached the problem lymphocyte detection as an instance segmentation problem by taking advantage of our own generated nuclei instance segmentation masks [117]. It also approves the reliability of the NuClick generated prediction masks, which is discussed in more details in the following subsection.

3.5.3 Segmentation Reliability Study

The important part of an interactive method for collecting segmentation is to see how the generated segmentation maps are reliable. To check the reliability of generated masks, we use them for training segmentation models. Then we can compare the performance of models trained on generated mask with the performance of models trained on the GTs. This experiment has been done for

Table 3.7: Results of segmentation reliability experiments.

	Result on MoNuSeg test set				Result on CPM test set			
	GT		$\text{NuClick}_{\text{CPM}}$		GT		$\text{NuClick}_{\text{MoNuSeg}}$	
	Dice	SQ	Dice	SQ	Dice	SQ	Dice	SQ
Unet	0.825	0.510	0.824	0.503	0.862	0.596	0.854	0.584
SegNet	0.849	0.531	0.842	0.527	0.889	0.644	0.881	0.632
FCN8	0.808	0.453	0.818	0.459	0.848	0.609	0.836	0.603

Table 3.8: Effect of disturbing click positions by amount of σ on NuClick outputs for nuclei and cells segmentation.

	Nuclei			Cells (WBCs)		
σ	AJI	Dice	PQ.	AJI	Dice	PQ.
1	0.834	0.912	0.838	0.954	0.983	0.958
3	0.834	0.911	0.837	0.954	0.983	0.958
5	0.832	0.911	0.835	0.953	0.983	0.957
10	0.821	0.903	0.822	0.953	0.982	0.957
20	-	-	-	0.950	0.979	0.955
50	-	-	-	0.935	0.961	0.943

nuclear segmentation task, where we trained three well-known segmentation networks (U-Net [47], SegNet [66], and FCN8 [29]) with GT and NuClick generated masks separately and evaluated the trained models on the validation set. Results of these experiments are reported in Table 3.7. Note that when we are evaluating the segmentation on MoNuSeg dataset, the NuClick model that generated the masks is trained on the CPM dataset. Therefore, in that case NuClick framework did not see any of MoNuSeg images during its training.

As shown in Table 3.7 there is a negligible difference between the metrics achieved by models trained on GT masks and the ones that trained on NuClick generated masks. Even for one instance, when testing on MoNuSeg dataset, Dice and SQ values resulted from FCN8 model trained on annotations of $\text{NuClick}_{\text{CPM}}$ are 0.01 and 0.006 (insignificantly) higher than the model trained on GT annotations, respectively. This might be due to more uniformity of the NuClick generated annotations, which eliminate the negative effect of inter annotator variations present in GT annotations. Therefore, the dense annotations generated by NuClick are reliable enough for using in practice. If we consider the cost of manual annotation, it is more efficient to use annotations obtained from NuClick to train models.

3.5.4 Sensitivity to Guiding Signals

Performance of an interactive segmentation algorithm highly depends on quality of the user input markers. In other words, an ideal interactive segmentation

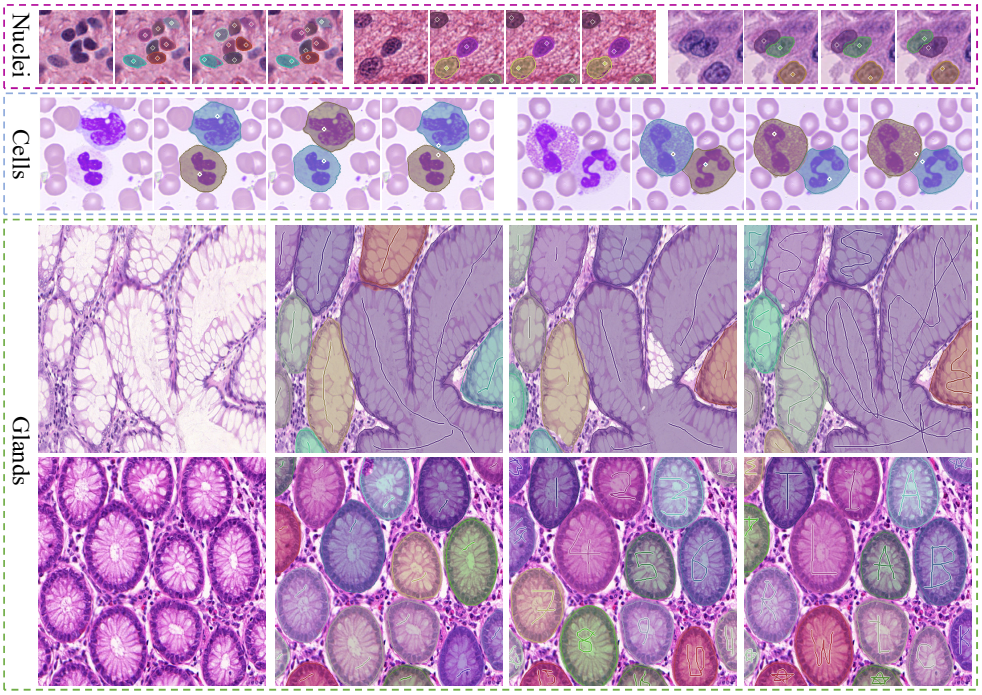

Figure 3.6: Example results of the NuClick, highlighting the variations in the user input. First and second rows show the predictions of the NuClick at different positions of clicks inside objects. The third and fourth rows demonstrate the predictions of the NuClick in presence of various shapes of squiggles. Solid stroke line around each object outlines the ground truth boundary for that object, overlaid transparent mask is the predicted segmentation region by the NuClick, and points or squiggles indicate the guiding signal for interactive segmentation. (Best viewed in color, zoom in to clearly see boundaries)

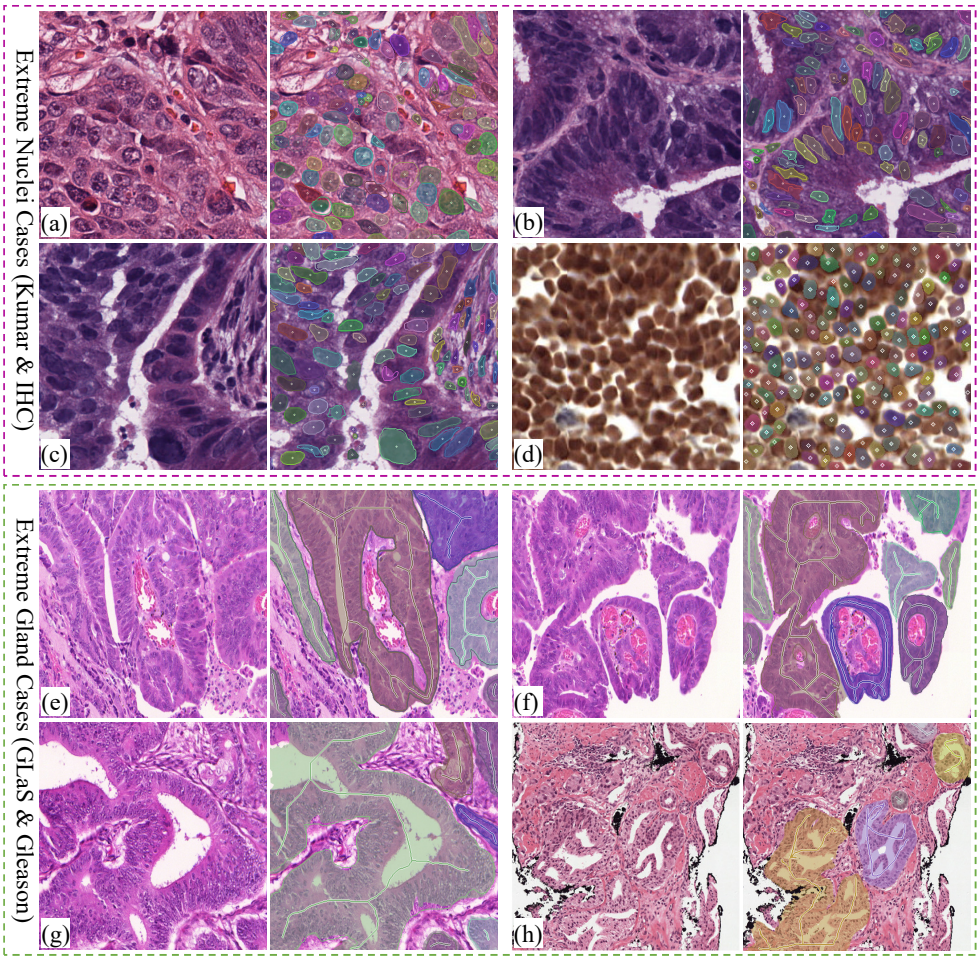

Figure 3.7: Extreme cases for nuclei and glands: clumped nuclei in H&E and IHC images (a-d) and irregular glands/tumor regions in cancerous colon and prostate images (e-h) are shown. In all images, solid stroke line around each object outlines the ground truth boundary for that object (except for d and e where the ground truth masks are unavailable), overlaid transparent mask is the predicted segmentation region by the NuClick, and points or squiggles indicate the provided guiding signal for interactive segmentation. (Best viewed in color, zoom in to clearly see boundaries)

tool must be robust against errors in the input annotations as much as possible. For instance, in nucleus or cell segmentation, an ideal segmentation tools should perform well to delineate boundaries of nuclei as long as user clicks fall inside the nuclei region i.e., the clicked point does not need to be located exactly at the center of the desired nuclei.

To assess the sensitivity of NuClick to the variations in the guiding signal, we design an experiment for nuclei and cell segmentation applications in which location of the guiding point in the inclusion map is perturbed by adding value of σ to the location of centroids. We repeat this experiment for different values of σ for both nuclei and cell segmentation applications and report the results in Table 3.8. For nuclear segmentation, jittering the location up to 10 pixels is investigated. It has been shown that disturbing the click position from the centroid up to 5 pixels does not considerably degrade the segmentation results. However, when the jittering amount is equal to $\sigma = 10$, all evaluation metrics drop by 1% or more. This reduction in metrics does not necessarily imply that NuClick is sensitive to click positions, because this fall in performance may be due to the fact that radius of some nuclei is less than 10 pixels and jittering the click position by 10 pixels cause it to fall outside the nuclei region therefore confusing the NuClick in correctly segmenting the desired small nucleus. However, even reduced metrics are still reliable in comparison with the resulted metrics from other methods as reported in Table 3.2.

The same trend can be seen for cell segmentation task in Table 3.8. However, for cells in our dataset we were able to increase the jittering range (up to 50 pixels) because in the WBC dataset, white blood cells have a diameter of at least 80 pixels. As one can see, the segmentation results are very robust against the applied distortion to the click position. Changing the click location by 50 pixels makes considerable drop in the performance which can be due to the same reason as we discussed for the nuclei i.e., amount of jittering is bigger than the average radius of some small cells.

Unfortunately, we can not quantitatively analyze the sensitivity of the NuClick to the squiggle changes, because its related changes are not easily measurable/paramtereizable. However, for two examples of histology images we showed the effect of changing the guiding squiggles on the resulting segmentation in Fig. 3.6. In this figure, the effect of changing the click position for two examples of nuclei segmentation and two examples of cell segmentation are also visualized. It is obvious from exemplars in Fig. 3.6 that NuClick successfully works with different shapes of squiggles as the guiding signal. Squiggles can be short in the middle or adjacent regions of the desired gland, or they can be long enough to cover the main diameter of the gland. They can be continuous curves covering all section and indentation of the gland geometry, or separated discrete lines that indicate different sections of a big gland. They can even have

arbitrary numerical or letters shape like the example in the last row of Fig. 3.6. In all cases, it is obvious that NuClick is quite robust against variations in the guiding signals which is due to the techniques that we have incorporated during training of the NuClick (randomizing the inclusion map).

It is worth mentioning that we have conducted experiments with training NuClick for gland segmentation using extreme points and polygons as guiding signals. Even with a considerable number of points on gland boundary or polygons with large number of vertices (filled or hollow), the network failed to converge during the training phase. However, we observed that even simple or small squiggles are able to provide enough guiding information for the model to converge fast.

We have also conducted another experiment to assess the sensitivity of NuClick on the exclusion maps. In other words, we want to see if eliminating the exclusion map has any effect on NuClick segmentation performance. To this end, we evaluate the performance of NuClick for nuclei segmentation on MoNuSeg dataset in the absence of exclusion map. Therefore in this situation the input to the network would have 4 channels (RGB plus inclusion map). The network is trained from scratch on the MoNuSeg training set with the new considerations and then evaluated on the MoNuSeg validation set. Results of this experiment are reported in Table 3.9. Based on Table 3.9, performance of the NuClick significantly drops when exclusion map is missing. That is because there are a lot of overlapping nuclei in this dataset and without having the exclusion map, the network has no clue of the neighboring nuclei when dealing with a nucleus that belongs to a nuclei clump.

3.5.5 Extreme Cases

To investigate the effectiveness of NuClick when dealing with extreme cases, output of NuClick for images with challenging objects (high grade cancer in different tissue types) are shown in Fig. 3.7. For example in Fig. 3.7a-c touching nuclei with unclear edges from patches of cancerous samples have been successfully segmented by NuClick. Additionally, Fig. 3.7d shows promising segmentation of densely clustered blood cells in a blurred IHC image from another domain (extracted from LYON19 dataset ([129])).

In Fig. 3.7e-f, images of glands with irregular shapes and their overlaid predictions are shown. As long as the squiggle covers the extend of gland, we can achieve a good segmentation. A noteworthy property of NuClick framework is its capability to segment objects with holes in them. In Fig. 3.7e-f, although margins of glands are very unclear and some glands have holes in their shape, NuClick can successfully recognizing boundaries of each gland. Further, if the squiggle encompass the hole, it will be excluded from final segmentation

Table 3.9: Performance of the NuClick on the MonuSeg dataset with and without exclusion map

	AJI	Dice	SQ	DQ	PQ
NuClick with ex. map	0.834	0.912	0.839	0.999	0.838
NuClick without ex. map	0.815	0.894	0.801	0.972	0.778

whereas if the squiggle covers part of holes in the middle of glands, they will be included in the segmentation. For instance, in Fig. 3.7g, a complex and relatively large gland is well delineated by the NuClick. Note that this gland contains a hole region which belongs to the gland and it is correctly segmented as part of the gland because the guiding signal covers that part. This is a powerful and very useful property that methods based on extreme points or bounding box like [104] and [97] do not offer.

We also show a cancerous prostate image (extracted from PANDA dataset ([131])) in Fig. 3.7h where the tumor regions are outlined by NuClick. Overall, these predictions shows the capability of NuClick in providing reasonable annotation in scenarios that are even challenging for humans to annotate. Note that for images in Fig. 3.7d,h the ground truth segmentation masks are not available, therefore they are not shown.

3.5.6 User Correction

In some cases, the output of models might not be correct, therefore there should be a possibility that user can modify wrong predictions. This is a matter of implementation of the interface in most cases, Hence, when the output is not as good as expected, the user can modify the supervisory signal by extending squiggles, changing the shape of squiggles or move the position of clicks. After the modification has been applied, the new modified supervisory signal is fed to the network to obtain new segmentation. This process is also briefly shown in Fig. 3.8.

3.6 Summary

In this chapter, we have presented NuClick, a CNN-based framework for interactive segmentation of objects in histology images. We proposed a simple and robust way to provide input from the user which minimizes human effort for obtaining dense annotations of nuclei, cell and glands in histology. We showed that our method is generizable enough to be used across different datasets and it can be used even for annotating objects from completely different data distributions. Applicability of NuClick has been shown across 6 datasets, where NuClick obtained state-of-the art performance in all scenarios. NuClick can

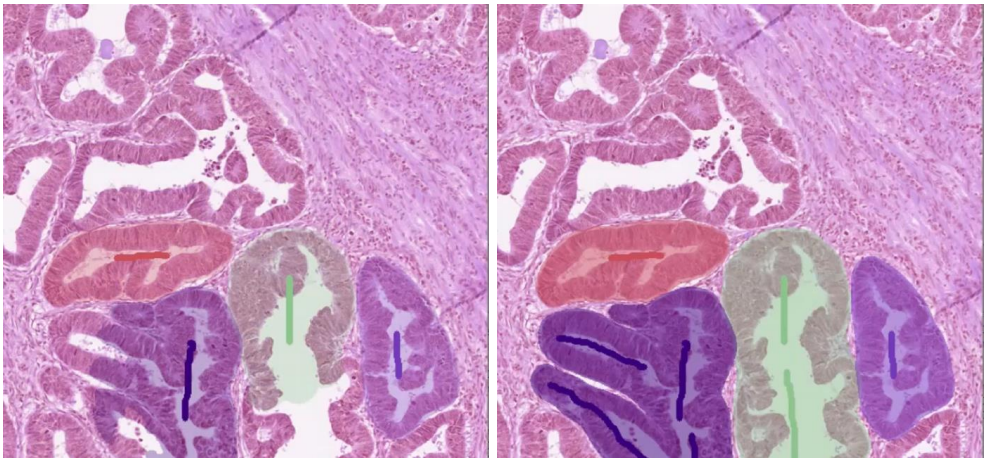

Figure 3.8: Segmentation process for gland: left image shows the initial marking and the initial results generated by the model. Since the strokes do not reflect the extend of glands, segmentation masks are not desirable. Therefore as shown in the right image user can add more strokes or modify the previous ones to achieve a good segmentation result.

also be used for segmenting other objects like nerves and vessels which are less complex and less heterogeneous compared to glands. We believe that NuClick can be used as a useful plug-in for whole slide annotation programs like ASAP [132] or Qupath [133] to ease the labeling process of the large-scale datasets. NuClick utilizes a fixed inference for prediction. One potential future direction can be considering the user feedback for upgrading gradient. In other words, the model upgrades itself as user annotates images and he/she is happy with annotation. NuClick has been developed for segmenting patches , the next step can be developing a platform for interacting with WSIs.

Chapter 4

Self-supervision for Classification of Pathology Images with Limited Annotations

4.1 Introduction

The recent surge in the area of computational pathology can be attributed to the increasing ubiquity of digital slide scanners and the consequent rapid rise in the amount of raw pixel data acquired by scanning of histology slides into digital whole-slide images (WSIs). These developments make the area of computational pathology ripe ground for deep neural network (DNN) models. In recent years, there have been notable successes in training DNNs for pathology image analysis and automated diagnosis of disease in the histopathology domain [134]. The performance and generalizability of most DNNs is, however, highly dependent on the availability of large and diverse amounts of annotated data. Although the use of digital slide scanners have made large amounts of raw data available, development of DNN based algorithms remains bottlenecked by the need for extensive annotations on diverse datasets.

In pathology, annotation burden can pose a large problem – even more so when compared to natural scene images. WSIs are by nature high resolution images (sometimes with slide dimensions as large as 200,000 $\times$ 150,000 pixels) – this hinders exhaustive annotations. For even simple use cases like detecting tumor regions or isolated tumor cells in WSIs, pathologists annotating the data need to look at regions of the tissue at multiple levels of magnification. So, even simple labeling of regions of interest can be quite demanding. This issue is compounded by the fact that the whole image can only be annotated

part by part owing to its large size. Further, the annotation effort requires expert domain knowledge and significant investment on the part of specialized pathologists. Overall, compared to annotating natural images, pathology images need experienced annotators whereas annotating natural images can be done by anyone, thresholds for meaningful performance in pathology is much higher than working with natural images and there are many complex cases in pathology where one expert can not make a decision. To overcome these challenges, when training DNNs on new pathology image datasets, it would be desirable to pursue one or both of the following strategies: (a) labeling small amounts of the new dataset and making use of the larger pool of the unlabeled data, and/or (b) using existing labeled datasets which closely match the new dataset.

For strategy (a), *semi-supervised deep learning* approaches that learn with small amounts of labeled data and leverage larger pools of unlabeled data to boost performance can be employed. These approaches have been widely demonstrated in the computer vision community for natural scene images. Particularly popular techniques include Mean Teacher [135] and Virtual Adversarial Training (VAT) [136]. Recently, these approaches have also been applied to the area of computational pathology to address tasks such as clustering [137], segmentation [138] and image retrieval [139]. However, due to the high dimensionality of the images, the multi-scale nature of the problem, the requirement of contextual information and texture-like nature of sub-patches extracted from slides, the direct translation of popular semi-supervised algorithms into pathology classification tasks is not feasible.

For strategy (b), *domain adaptation* approaches that transfer knowledge from existing resources for related tasks to the classification task-at-hand can be employed. However, due to variations in tissue, tumor types, and stain appearance during image acquisition, different pathology image datasets appear quite distinct from one another. In addition, for some rare tissue or tumor types, there may be no annotated datasets available for such knowledge transfer. Hence, direct translation of existing domain adaptation algorithms which work for natural vision images may not be possible. Yet, unlabeled data for related tasks are largely available and are less prone to bias [140]. Hence, when dealing with limited annotations, such unlabeled data can be used to capture the shared knowledge or to learn representations that can improve model performance.

To address the dual challenges of low annotations and domain adaptation in histopathology, it is possible to use unlabeled data in a self-supervised manner. In this setup, the model is supervised by labels that come inherently from the data itself without any additional manual annotations. These labels can represent distinct morphological, geometrical and contextual content of the

images. Models trained on these 'free' labels can learn representations that can improve performance for a variety of tasks such as classification, segmentation and detection [141]. Self-supervision tasks can be used together with the main supervised task in a multi-task setup to improve performance for semi-supervised learning and domain adaptation [142]. However, self-supervised tasks proposed in the literature so far are mainly based on characteristics of natural scene images, which are very different from histology images. For instance, common self-supervision tasks focus on predicting the degree of rotation, flipping, and/or the relative position of objects. While these are meaningful concepts for natural scene images, they do not carry much relevance for histopathology images. Specifically, while the degree of rotation could help to also learn semantic information present in a natural image, it would not make sense for pathology images because they have no sense of global orientation [143].

In this chapter, we propose the **Self-Path**[1] framework to leverage self-supervised tasks customized to the requirements of the histopathology domain, and enhance DNN training in scenarios with limited or no annotated data for the task at hand. Our main contributions are summarized as follows:

- We introduce a generic and flexible self-supervision based framework, Self-Path, for classification of pathology images in the context of limited or no annotations.
- We propose 3 novel pathology pathology specific self-supervision tasks, namely, prediction of magnification level, solving the magnification jigsaw puzzle and prediction of the Hematoxylin channel, aimed at utilizing contextual, multi-resolution and semantic features in histopathology images.
- We conduct a detailed investigation on the effect of various self-supervision tasks for semi-supervised learning and domain adaptation for three datasets.
- We demonstrate that Self-Path achieves state-of-the art performance in limited annotation regime (when 1-2% of the whole dataset is annotated) or even when no annotations are available (in the case of domain adaptation).

4.1.1 Related Work

Semi-supervised Learning: Semi-supervised deep learning approaches are widely studied in the computer vision literature [144]. Popular methods utilize

[1] Code as available at: https://github.com/navidstuv/self$_path/tree/upload$

forms of pseudo labelling and consistency regularization, and utilize small amounts of labeled data alongside larger pools of unlabeled data for learning. Pseudo-labeling approaches [145] use available labels to train a model and impute labels on the unlabeled samples which are in turn used in training. MixMatch extends pseudo-labeling by adding temperate sharpening along with the mix-up augmentation [146] . Consistency-based methods regularize the model by ensuring stable outputs for various augmentations of the same sample. These can be done by enforcing consensus between temporal ensembles of network outputs like in Pi-Model [147], or between perturbed images fed to a network and its EMA averaged counterpart like in Mean Teacher[135]. Virtual adversarial training(VAT) [136] generates the perturbed images in an adversarial fashion to smooth the margin in the direction of maximum vulnerability. These methods ensure generalizability against significant image perturbations, move the margin away from high-density regions, and enable strong performance on benchmark natural scene image tasks with low annotation budgets.

However, semi-supervised learning has not been sufficiently explored in pathology image analysis. At the time of this writing, only 6 papers investigate semi-supervised learning for the histopathology domain. In [138], Li et. al proposed an EM-based approach for semi-supervised segmentation of histology images. [137] proposed a cluster based semi-supervised approach to identify high-density regions in the data space which were then used by supervised SVM in finding the decision boundary. Jaiswal *et al.* [148] used pseudo-labels for improving the network performance for metastasis detection of breast cancer. Su *et al.* [149] employed global and local consistency losses for mean teacher approach for nuclear classification. Shaw et. al [150] also proposed to use pseudo-labels of unlabeled images for fine-tuning the model iteratively to improve performance for colorectal image classification. Deep multiple instance learning and contrastive predictive coding were used together in [151] to overcome the scarcity of labeled data for breast cancer classification. Yet, there is scope for improvement to close the gap between fully supervised baselines and semi-supervised methods employing just a few labeled pathology images.

Domain Adaptation: Domain adaptation methods focus on adapting models trained on a source dataset to perform well on a target dataset. Leading-edge techniques mainly use adversarial training for aligning the feature distributions of different domains. Popular domain-adversarial learning-based methods[152, 153] use a domain discriminator to classify the domain of images. These methods play a minimax game where the discriminator is trained to distinguish the features from the source or target sample, while the feature generator is trained to confuse the discriminator. [154] employed adversarial learning

and minimized Wassertein distance between domains to learn domain-invariant features. Image-translation methods minimize the discrepancy between the two domains at an image-level [155]. In pathology, Ren *et al.* [156] employed adversarial training for domain adaptation across acquisition devices (scanners) in a prostate cancer image classification task. [157] used CycleGAN to translate across domains for a cell/nuclei detection task. [158] introduced a measure for evaluating distance between domains to enhance the ability to identify out-of-distribution samples in a tumor classification task. Yet, most practical domain adaptation techniques require labeling of target domain data, and the applicability of state-of-the-art unsupervised domain adaptation approaches for histopathology is yet to be widely established.

Self-Supervision: Self-supervision employs pretext tasks (based on annotations that are inherent to the input data) to learn representations that can enhance performance for the downstream task [141]. Autoencoders [159] are the simplest self-supervised task, where the goal is to minimize reconstruction error and the proxy labels are the values of image pixels. Other self-supervised tasks in the literature are image generation [141], inpainting [160], colorizing grayscale images [161], predicting rotation [162], solving jigsaw puzzle [163], and contrastive predictive coding [164]. Perhaps the main difference between contrastive learning approaches and methods like ours is that while our method caters to a specific use case domain and the task at hand is to come up with self-supervision tasks, the contrastive learning approaches offer the advantage of a more generic framework for learning representations potentially at the cost of losing performance in a very specific use case domain (such as histopathology). Although the classical self-supervision approaches requires no additional annotations, it is also possible to leverage small amounts of labeled data within a self-supervision framework. For example, S4L [142] showed that the pretext task (e.g., rotation, self-supervised exemplar [165]) can benefit from small amount of labeled data alongside larger unlabeled data. Moreover, some works [166, 167] demonstrated the effect of self supervised tasks for domain adaptation, where in [167] the effect of various self-supervised tasks have been shown for domain alignment. Particularly, solving jigsaw puzzle [167] has been proved to be a beneficial pretext task for domain generalization.

As there is no large labeled dataset akin to ImageNet for pretraining in the pathology domain, self supervised learning offers potential to obtain pre-trained model that preserves the useful information about data in itself. Although one recent study [168] explored self-supervised similarity learning for pathology image retrieval, much of the self-supervision literature is focused on computer vision applications. A key challenge in applying self-supervision to pathology-specific applications is to define the pretext task that will be most beneficial. As such, systematic analysis and derivation of pretext tasks customized for a

range of histopathology applications would be desirable.

4.2 Problem Formulation

We now define the problem of semi-supervised learning and domain adaptation for pathology image classification. Consider a whole slide image (WSI) that is comprised of a number of disjoint or overlapping 'patches'. We denote an input image or 'patch' as $\mathbf{x}$ and its associated class label as y.

4.2.0.0.1 Semi-supervised Learning

We consider a set of n_l limited labeled images $S_L = \{(\mathbf{x}_i^l, y_i)\}_{i=1l}^{n}$, and a set of $m_l >> n_l$ unlabeled images $S_U = \{(\mathbf{x}_i^u)\}_{i=1l}^{m}$. The semi-supervised framework seeks to leverage the large pool of unlabeled images in S_U to enhance the generalizability of learning with fewer labeled images in S_L. Generally, in the semi-supervised setting, both S_L and S_U are from the same distribution.

4.2.0.0.2 Domain Adaptation

We define a source domain S comprising a set of η_s labeled images $D_s = \{\mathbf{x}_i^s, y_i^s\}_{i=1}^{\eta_s}$. Likewise, we have a target domain T comprising a set of η_t unlabeled images $D_t = \{\mathbf{x}_i^t\}_{i=1}^{\eta_t}$. Both source and target domains have the same labels. Further, source and target domains have related task characteristics, but their data distributions are distinct.

4.3 Methods

Our proposed Self-Path framework is depicted in Figure 4.1. To address label scarcity for the main classification task (main task), Self-Path leverages self-supervision and informs the supervised learning for the main task with the self-supervised learning for pretext tasks. Further, our proposed framework employs a multi-task learning approach to learn class-discriminative and domain-invariant features that would generalize with limited annotated data. Specifically, Self-Path (a) can leverage one or more pathology-specific or pathology-agnostic pretext tasks, (b) is amenable to adversarial or non-adversarial training, and (c) allows flexibility to incorporate semi-supervised, generative learning and/or domain adaptation approaches. We now formally describe the multi-task learning objective and detail the pretext tasks that are used along with the main task.

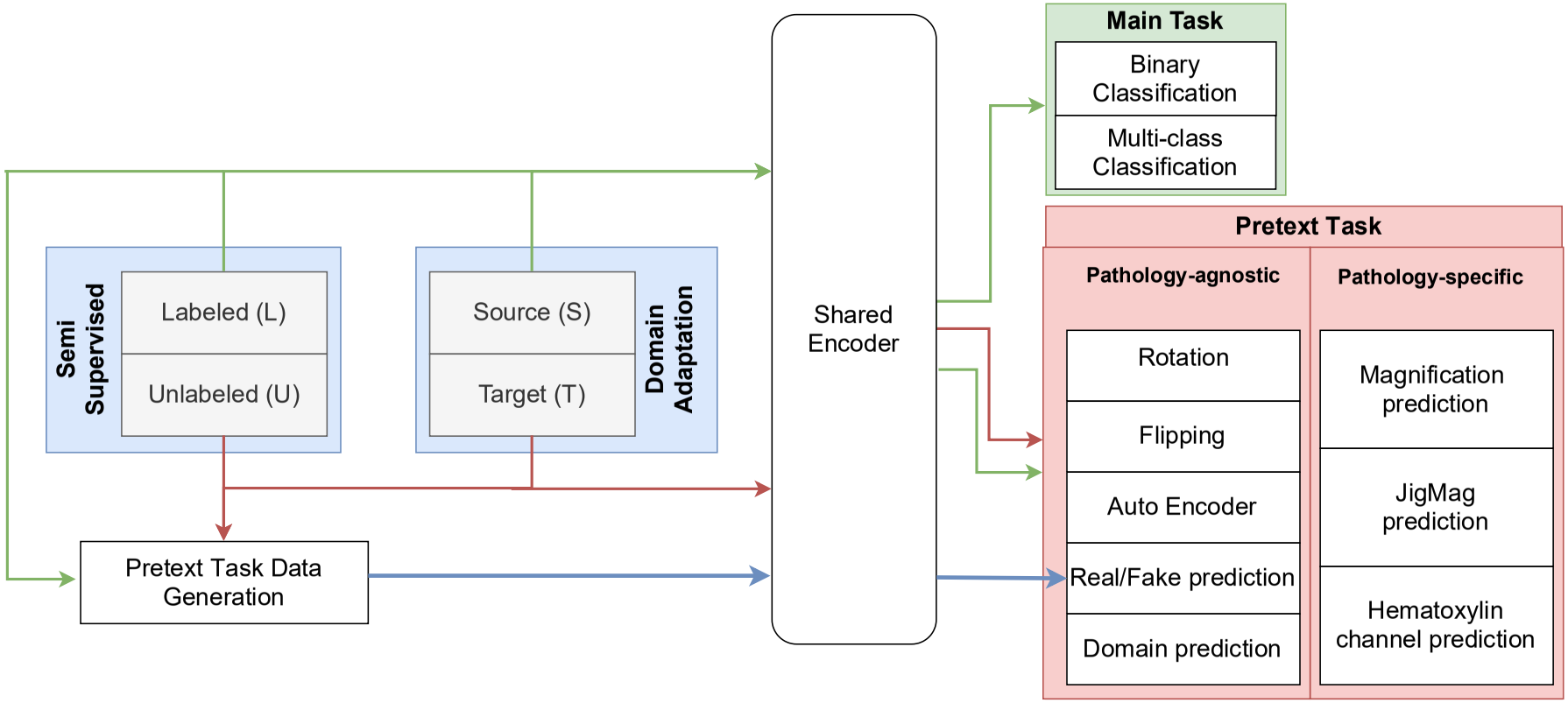

Figure 4.1: Overview of Self-Path : The framework employs self-supervised pretext tasks. Pretext tasks can be added atop a shared encoder to learn useful representations and enhance semi-supervised learning or domain-adaptation. Green, red and blue lines indicate the flow of labeled, unlabeled and generated images, respectively. Generated images are used only for the generative task.

4.3.1 Multi-task Learning

Our proposed approach trains the model using the main and pretext tasks in conjunction. The framework comprises a shared encoder which learns features that are common to both the pretext task and the main task. Each task usually has a separate head connected to the shared encoder and learning for all tasks is optimized simultaneously. Formally,

$$\begin{aligned} \underset{\theta_c,\theta_e,\,\theta_{p_1},..\theta_{p_k}}{\operatorname{argmin}} & \frac{1}{n_l}\sum_{i}^{n_l} L_c(F_c^{\theta_c}(F_e^{\theta_e}(\mathbf{x}_i^l)),\ y_i) \\ +\frac{1}{n_l}\sum_{k=1}^{K}\alpha_{p_k}\sum_{i}^{n^l} & L_{p_k}(F_{p_k}^{\theta_{p_k}}(F_e^{\theta_e}(\mathbf{x}_i^l)),\ r_{ik}^l) \quad , \\ +\frac{1}{n_u}\sum_{k=1}^{K}\alpha_{p_k}\sum_{i}^{n_u} & L_{p_k}(F_{p_k}^{\theta_{p_k}}(F_e^{\theta_e}(\mathbf{x}_i^u)),\ r_{ik}^u) \end{aligned} \tag{4.1}$$

where K is the number of pretext tasks, r is the label for pretext task; L_c and L_{p_k} are the losses for the main and pretext tasks, respectively; F_e, is the shared encoder, F_c is the function for main task and F_{p_k} is the function of kth pretext task; θ_c, θ_e and θ_{p_n} are parameters of main task classifier, shared encoder and pretext tasks, respectively; α_{p_k} indicates weights for different tasks; and n_l and n_u indicate the number of labeled and unlabeled images, respectively. When this model is used for semi-supervised learning, the labeled and unlabeled images come from the same domain. When used for domain adaptation, the labeled images come from source domain and unlabeled images come from the target domain.

4.3.2 Self-Supervision

The self-supervision utilizes one or more pretext tasks to leverage information in the unlabeled images and improve performance for the main task. Our setup employs both pathology-specific and pathology-agnostic self-supervised tasks. Every pretext task p_k is defined by a transformation function g_k applied to input x, and an implicit label r_k for the transformed input $\tilde{x} = g_k(x)$. Then, the objective function L_{p_k} is the objective for learning the self-supervised classification task that maps $\tilde{x}$ to r_k.

4.3.3 Pathology-specific Pretext tasks for Self-supervision

Histopathology images can vary in shape, morphology and arrangement of the nuclei across tissue types and disease conditions. Learning these features or semantic representations of these features can enable generalizable classification models that can more effectively transfer knowledge across domains. Therefore, we design pathology-specific pretext tasks that cater to morphology, context and shapes of nuclei as detailed below :

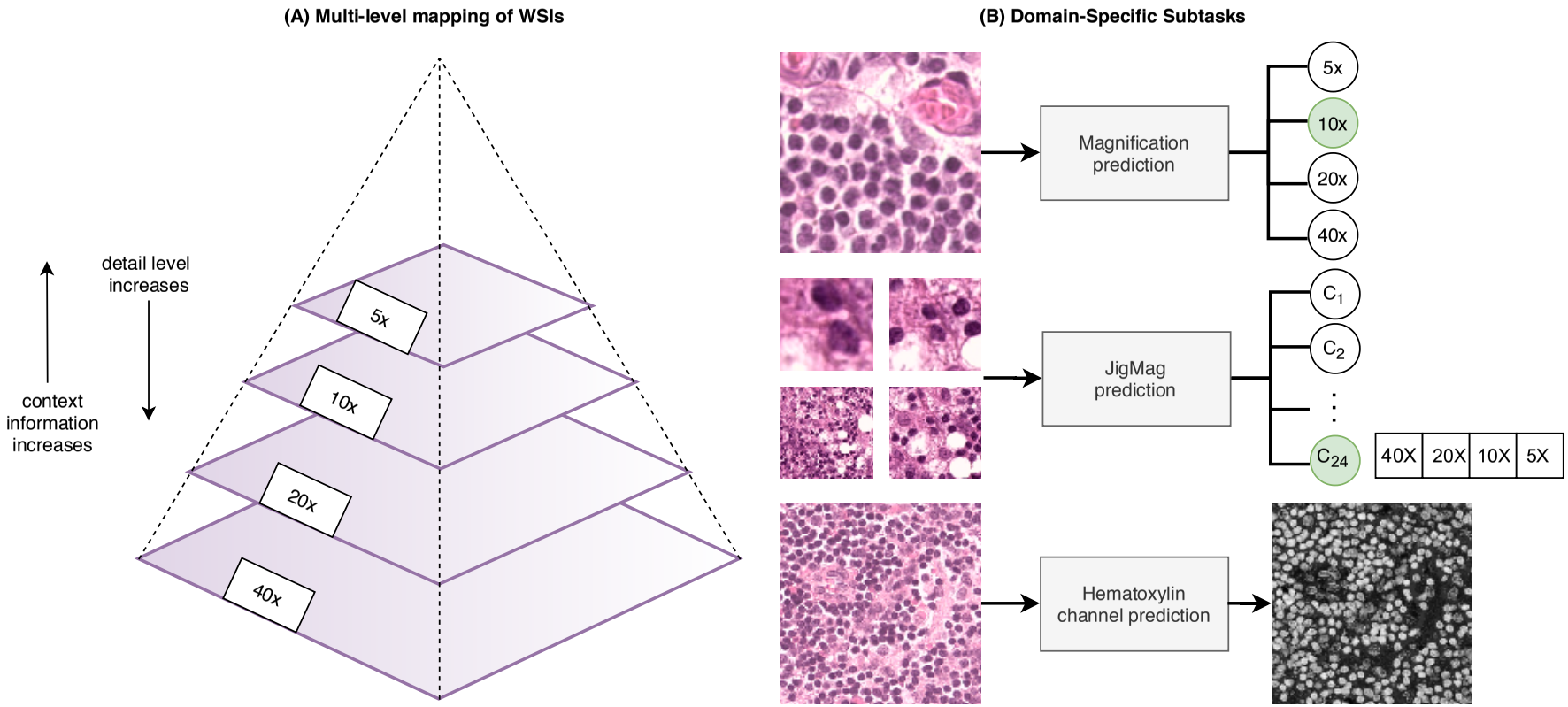

Figure 4.2: (A) Whole slide images (WSI) in pathology slides organized hierarchically - each level trades-off the degree of detail against the availability of contextual information. (B) Pathology specific pretext tasks created for Self-Path.

4.3.3.1 Magnification Prediction

Histopathology images are often generated and viewed at various standard magnification levels. Considering an image of fixed size, higher magnifications provide more details but less context, whereas lower magnifications allow less details but more context of tissue region. Pathologists assessing an image tend to infer important semantic information by iterating between detail and context – i.e., by zooming in and out on WSIs or by looking at different magnification levels [2]. In other words, magnification levels are implicitly correlated with important semantic information. Therefore, to enable the classification model to learn semantic information, we set up a pretext task focused on estimating magnification level of the image. Specifically, the pretext task focuses on classifying the input image to 1 of 4 magnification levels (40×, 20×, 10× and 5×). We extract images or patches from WSIs at these magnification levels Figure 4.2 (A). If a magnification level is not available, we obtain the patches by (bi-linear) resizing patches from other magnification levels that are available. For example, to obtain 128 × 128 patches at 5×, we extract patches of 1024 × 1024 at 40× and down-sample by factor of 4. We then feed the extracted images to the network, which learns by minimizing a cross-entropy objective function.

4.3.3.2 Solving Magnification Puzzle (JigMag)

A basic problem in pattern recognition is the jigsaw task of retrieving an original image from its shuffled parts [169]. Convolutional neural networks (CNNs) have been employed to solve the jigsaw puzzle [140]. To solve the jigsaw puzzle, it is known that the network should learn the global semantic representation of images. This is achieved by concentrating on the differences between tiles and their positions while avoiding low level statistics [140]. In histopathology, objects are smaller compared to natural scene images, and there is no specific ordering among the objects. For example, the relative positions of different parts of dog in a natural scene image is consistent, however we do not have a similar concept in histopathology. Therefore, solving the jigsaw puzzle is by itself not sufficient for learning useful semantic representations in histopathology.

Instead, we propose to create a puzzle to reflect the magnification and context characteristics of histopathology images. Conceptually, classification can be enhanced by having the network implicitly learn object size and associated contextual information. Hence, we propose a pretext task focused

[2] Magnification levels and their corresponding resolutions vary for each scanner. However by observing one particular magnification of an image, other magnifications can be perceived easily for the same scanner.

on solving this magnification and context puzzle. In this puzzle, an image consists of image tiles with various magnifications and the network is tasked with predicting their arrangement. This set up caters also to the need to classify images containing objects with varying shapes and sizes.

Specifically, we define v as a vector of image orders in a 2×2 grid where each grid includes a specific magnification. For example $v = [0, 1, 2, 3]$ defines that image with magnification 5× is on top left corner, 10× is on top right and so on. We consider 24 different orders of magnification. To construct our proposed jigsaw puzzle, we first extract patches of size 512×512 at 40× magnification then each part of the puzzle is constructed by down-sampling and or center-cropping to the size of 64×64, where each reflects specific context and resolution of the the original extracted patch. This pretext task employs a cross entropy loss function.

4.3.3.3 Hematoxylin Channel Prediction

Commonly, histopathology images are stained with Hematoxylin and Eosin (H&E). In H&E images, hematoxylin turns the palish color of nuclei to blue and eosin changes the color of other contents to pink. Color deconvolution methods have been applied to specifically identify cell nuclei in H&E images. Therefore by extracting hematoxylin channel, one can locate the nuclei and their approximate shape. Pathologists often use the location, shape and morphology of nuclei in the hematoxylin channel to diagnose or classify histopathology images (especially for malignant features).

Therefore, one way to enhance learning of useful representations is to enable the classifier to identify the nuclei and their associated characteristics. We choose to define a pretext task focused on predicting the hematoxylin channel from H&E. We use the approach in [170] to extract the hematoxylin channel in our images and define the ground truth for the self-supervision task. We scale the values of hematoxylin channel in the range [0,1] and employ a mean absolute loss for optimizing this task.

4.3.4 Pathology-agnostic Self-supervision Tasks

The literature has investigated various pretext tasks like rotation prediction, flipping, image reconstruction [141, 162]. These were however, not tailored for pathology data. Here, we systematically study and benchmark efficacy of these pretext tasks for semi-supervised learning and domain adaptation in histopathology applications.

4.3.4.1 Prediction of Image Rotation

For predicting rotation, the input image is rotated with degrees of 0°, 90°, 180° and 270° corresponding to the labels 0, 1, 2 and 3, respectively [162].

4.3.4.2 Prediction of Image Flipping

The label assigned to the horizontal flipping of image is 1 and 0 if not flipped.

4.3.4.3 Image Reconstruction with Autoencoder

For reconstructing the image, a convolutional decoder is used on top of the feature extractor [159], similar to one for predicting hematoxylin channel however 3 channels is considered for output.

4.3.4.4 Real vs Fake Prediction (Generative)

The generative learning literature has shown that predicting whether an image is real or fake can help to learn useful representations for classification [171]. Therefore, we introduce a generative pretext task focused on real vs. fake prediction. To learn this pretext task, we train a generative network in an adversarial fashion by using unlabeled samples. While one could use a shared encoder to extract features, we found that it is easier to employ a simpler encoder/discriminator similar to the generative adversarial network (GAN) in [171].

Formally, real images are drawn from distribution D_{real}, and the generative function learns the distribution D_{gen} where the goal is to align this two distributions ($D_{gen} \sim D_{real}$). The generator $G(.)$ takes predefined noise variables z from a uniform distribution D_{noise}. The objective function is defined as:

$$\begin{aligned} L_{dis} &= -\mathrm{E}_{x\sim\, D_{real}}[\log[1 - F_{Dis}(F_e(x))]] \\ &-\mathrm{E}_{x\sim\, D_{gen}}[\log[F_{Dis}(F_e(x))]] \qquad , \qquad (4.2) \\ L_{gen} &= \|\mathrm{E}_{x\sim D_{real}}\,[F_e(x)] - \mathrm{E}_{z\sim D_{noise}}\,[F_e(G(z))]\|_1 \end{aligned}$$

where L_{gen} and L_{dis} are the generator and discriminator losses, respectively. $F_e(x)$ is the feature from intermediate layer of feature extractor (last layer before fully connected layers) and $F_{Dis}(F_e(x))$ is the output of the discriminator (fake/real head).

4.3.4.5 Domain Prediction

In order to learn useful representations to facilitate domain adaptation, it is useful to have a network learn the common features between source and target domains. Therefore, we introduce a pretext task to predict if the image belongs to source or target domain, and employ it in combination with other pretext tasks for the domain adaptation experiments.

For this pretext task, we employ a domain adversarial neural network (DANN) [153]. DANN includes a minimax game where discriminator H_d (domain prediction head) is trained to distinguish between the source and target domain, and the feature extractor is simultaneously trained to confuse the discriminator. Therefore, to extract the common or domain-invariant features, the parameters of feature extractor θ_e (shared encoder in the multi-task setup) are learned by maximizing the loss of domain discriminator L_d, while parameters of the domain discriminator are learned by minimizing the loss of domain discriminator. Parameters of the main task F_c are also minimized to ensure good performance on the main task. Formally:

$$\begin{aligned} \underset{\theta_c,\theta_e}{\operatorname{argmin}}\underset{\theta_d}{\max} \frac{1}{\eta_s} \sum_{i=0}^{\eta_s} L_c(F_c^{\theta_c}(F_e^{\theta_e}(\mathbf{x}_i^s)),\ y_i) + \\ -\frac{\alpha_d}{\eta_s+\eta_t}(\sum_{i=1}^{\eta_s+\eta_t} L_d(F_d^{\theta_d}(F_e^{\theta_e}(\mathbf{x}_i)),\ d_i) \end{aligned} , \tag{4.3}$$

where d_i is the domain label for $\mathbf{x}_i$ and α_d is a coefficient for discriminator loss. In practice, we apply domain confusion using the Gradient Reversal Layer (GRL), where the gradients of L_d with respect to the gradients of feature extractor parameters θ_e ($\frac{\partial L_d}{\partial \theta_e}$) are reversed during back-propagation.

4.4 Experiments

4.4.1 Datasets

4.4.1.1 Camelyon16

We used the Camelyon 16 challenge dataset [15] that contains 399 H&E stained WSIs obtained on patients with breast cancer metastasis in the lymph nodes. The WSIs were acquired from 2 different centers, namely: Radboud University Medical Center (RUMC) and University Medical Center Utrecht (UMCU). RUMC images were generated by a digital slide scanner (Pannoramic 250 Flash ; 3DHISTECH) with a 20× objective lens (0.243 μm × 0.243 μm) and UMCU images were produced using a digital slide scanner (NanoZoomer-XR Digital slide scanner C12000-01; Hamamatsu Photonics) with a 40× objective lens (0.226μm × 0.226 μm). The tumor regions are exhaustively annotated by pathologists. We used the official training and testing splits comprising 270

and 129 WSIs, respectively. We randomly sampled 34 WSIs of the training set for validation. For our experiments, we randomly extracted patches from both normal and tumor regions (Table 4.1).

4.4.1.2 LNM-OSCC

LNM-OSCC is an in-house dataset comprising 217 H&E WSIs obtained on patients with Oral Squamous Cell Carcinoma (OSCC). Of these 217 patients, 140 have metastases in the cervical lymph nodes and 77 do not manifest metastases in the cervical lymph nodes. The WSIs were acquired from 2 hospitals using 2 different scanners – (a) 98 WSIs scanned with 40× objective lens using IntelliSite Ultra Fast Scanner (0.25 μm/pixel) at University Hospital Conventry and Warwickshire (UHCW), and (b) 119 WSIs scanned at the School of Medical Dentistry in Sheffield University by Aperio/Leica CS2 with 20× objective lens (0.2467 μm/pixel). The training set comprises 100 WSIs, the validation set 14 WSIs and testing set 103 WSIs. For those cases in the training and validation sets that have metastases, a sampling of the tumor and normal regions were delineated with bounding box annotations by pathologists. For the testing set, the tumor regions were exhaustively annotated at the pixel-level.

4.4.1.3 Kather

This dataset contains 107,180 image patches from H&E stained WSIs comprising human colorectal cancer (CRC) and normal tissue. For this dataset, only patches were available (no WSIs).The dataset covers 9 tissue classes: Adipose (ADI), background (BACK), debris (DEB), lymphocytes (LYM), mucus (MUC), smooth muscle (MUS), normal colon mucosa (NORM), cancer-associated stroma (STR), colorectal adenocarcinoma epithelium (TUM). We used the official data splits comprising 100k patches for training and 7180 patches for testing. We randomly sampled 20k patches of the training set for validation.

4.4.2 Data Summary

Figure 4.3 shows some illustrative examples of the different datasets used in our study. The overall data statistics are shown in Table 4.1. For Camleyon16 and LNM-OSCC datasets, we extracted patches from the WSIs, and patches are distributed equally for each class. For our main task the patch extraction size is 128 × 128 at 10×. The Kather dataset patches are sized 224 × 224 and we resized to 128 × 128 for our experiments.

[!t]

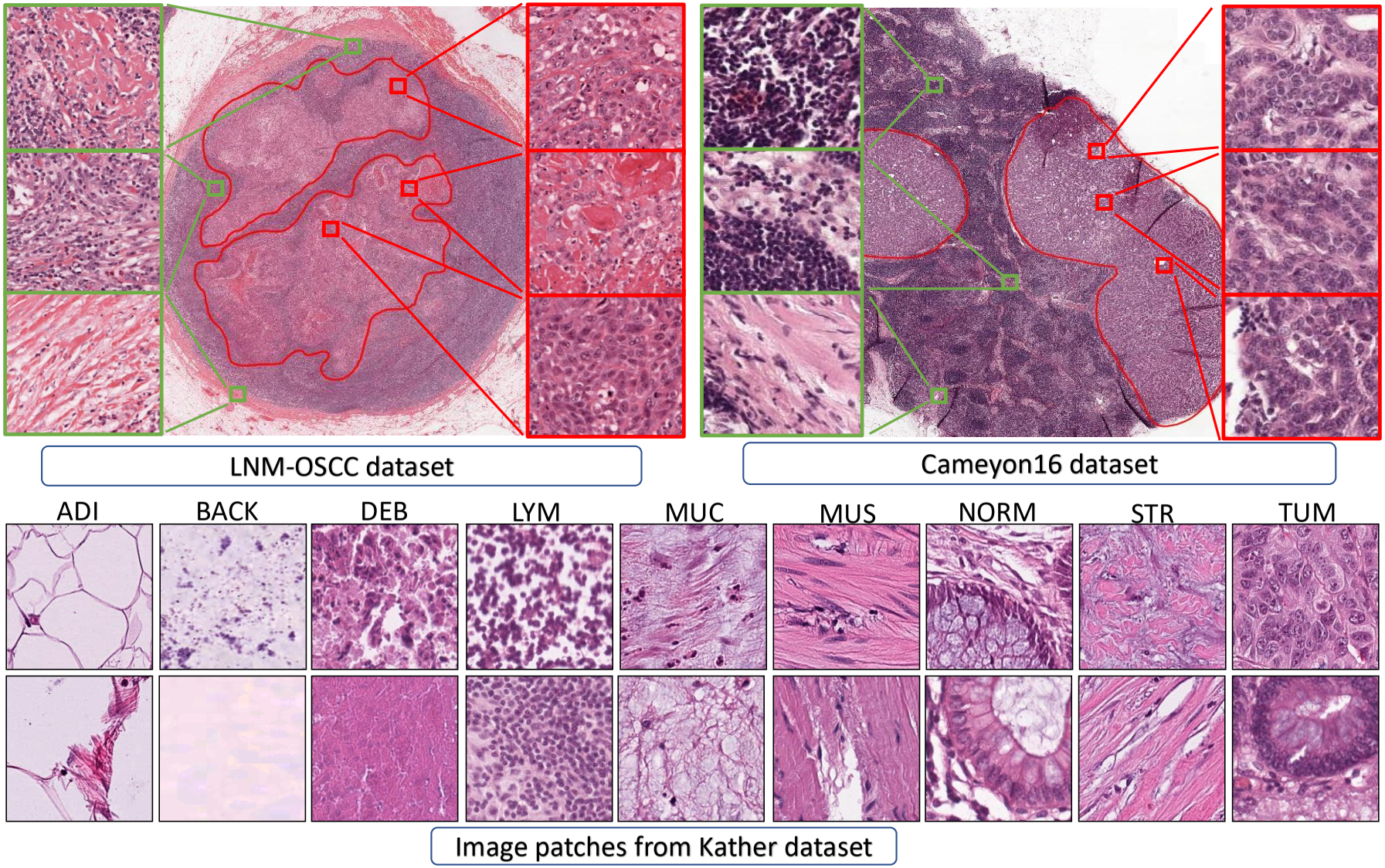

Figure 4.3: Exemplar images of different datasets that are used in this study. Red and green boxes denote the tumor and normal image patches.

Table 4.1: Number of WSIs and patches in each dataset.

		Train	Validation	Test
Camelyon16	WSIs	236	34	129
	patches	67054	15586	16562
LNM-OSCC	WSIs	100	14	103
	patches	55416	7224	14472
Kather	patches	79994	20006	7180

4.4.3 Experimental Setup

4.4.3.1 Networks

We chose Resnet50 [56] as the feature extraction backbone for all our experiments. The classifier head consists of adaptive average pooling which is followed by fully connected layer and softmax. The decoder head for reconstructing image and predicting hematoxylin channel is similar to the UNet decoder [47] (Table A.3) without using any skip connections. While using the real vs fake pretext task for image generation, we utilize the architecture presented in [171] (Table A.2) and find that this simpler feature extractor allows easy and robust convergence for the image generator.

4.4.3.2 Implementation Details

When Resnet50 is used as the shared encoder, we trained the network for 200 epochs. Our experiments used batch size 64, Adam optimizer, and learning rate of 10^{-3}. We fed batches of labeled and unlabeled images to the network separately. Therefore an epoch is defined as one full step through all the unlabeled images. Since our self-supervised experiments utilize fewer labeled images than unlabeled images, the labeled images are repeated in an epoch. Experiments related to real vs fake prediction used number of epochs and batch size of 500 and 32, respectively; and employed Adam optimizer with learning rate of 3×10^{-4}. For training model in multitask setup, we separately input batches of images for each task to the network and then sum their losses with their corresponding weights. Finally we back-propagate the whole loss through the network.

4.4.4 Results of Semi-Supervised Experiments

Here, we compare the effect of different self-supervision tasks for semi-supervised learning. We compare our models against the popular semi-supervised benchmarks, namely Mean Teacher [135] and VAT [136]. We also compare with teacher-student chain [150] (TSchain). TSchain is a recent semi-supervised approach for histopathology domain, that predicts the pseudo-labels for the

unlabeled data and then uses all images for iteratively retraining the model. For performance evaluations, we follow the typical protocol of varying the annotation budget for the training set while maintaining a fixed validation set, and reporting AUCs (average across 3 seeds) on the test set.

4.4.4.1 Results for LNM-OSCC Dataset

We report performance of each of the self-supervised tasks on LNM-OSCC dataset in Table 4.2. We have evaluated the model performance in terms of AUROC (Area Under the Receiver Operating Characteristic) for different annotation budgets (1%, 4%, 5%, 10% and 20% of the available WSIs). The semi-supervised approaches train on a combination of the labeled and unlabeled WSIs. The supervised baseline is only trained on labeled images without utilizing any unlabeled images.

We observe from Table 4.2 that at very low annotation budgets, pathology specific self-supervised tasks outperform the baselines and the pathology agnostic self-supervised tasks. For instance, at annotation budgets of 1% (1 labeled WSI, 134 labeled patches) and 4% (4 labeled WSIs, 1120 labeled patches), JigMag task has the best performance. At annotation budgets of 1% and 2%, Hematoxylin and magnification tasks outperform pathology agnostic tasks and generative tasks. When annotation budget increases to 10%, we observe that the generative task performs much better (AUC 95.4%), suggesting that the generated images can help the classifier to boost the performance. Overall, our LNM-OSCC experiments suggest that for limited annotation budgets, pathology specific pretext tasks are helpful for enhancing the model performance, with JigMag outperforming other approaches.

4.4.4.2 Results for Camelyon16 Dataset

We report performance of each of the self-supervised tasks on Camelyon16 dataset in Table 4.3. We have evaluated the model performance in terms of AUROC (Area Under the Receiver Operating Characteristic) for different annotation budgets (1%, 2%, 5%, 10% and 20% of the available WSIs). The semi-supervised approaches train on a combination of the labeled and unlabeled WSIs. The supervised baseline is only trained on labeled images without utilizing any unlabeled images.

Similar to LNM-OSCC dataset, pathology specific tasks outperform other semi supervised methods. In particular, the JigMag task improves the performance over the supervised baseline by 13.4%, 11.8% and 6.2% at 1% (2 WSIs), 2% (4 WSIs) and 5% (8 WSIs) annotation budgets, respectively. At 1% annotation budget, only magnification and JigMag outperform mean teacher and supervised baseline. Unlike LNM-OSCC, the generative model cannot

Table 4.2: LNM-OSCC Results for Different Annotation Budgets. Annotation budget is defined as the percentage of available WSIs that are labeled. The number of patches associated with each budget are indicated in the parentheses. The supervised upper bound performance when using all labeled data is 98.4%.

% Labeled WSIs (No. Patches)	1%(134)	2%(1024)	5%(1880)	10%(3334)	20%(7558)
	AUROC(%)	AUROC(%)	AUROC(%)	AUROC(%)	AUROC(%)
	Baselines				
supervised baseline	73.4 ± 2.0	76.1 ± 5.3	85.3 ± 6.3	86.3 ± 2.7	96.3 ± 0.3
mean teacher [135]	75.1 ± 4.5	78.4 ± 5.6	86.2 ± 7.6	91.4 ± 1.2	**97.4 ± 0.3**
VAT [136]	74.5 ± 5.6	77.4 ± 3.3	85.3 ± 4.3	92.1 ± 1.2	96.5 ± 0.9
TS chain [150]	75.3 ± 2.4	79.3 ± 2.5	85.2 ± 3.1	94.1 ± 1.7	97.2 ± 0.2
	Pathology-Agnostic Self-supervised Tasks				
rotation	74.5 ± 5.6	76.3 ± 4.2	88.4 ± 1.5	93.2 ± 0.3	96.2 ± 0.1
flipping	74.6 ± 4.0	74.2 ± 5.3	85.3 ± 4.1	91.4 ± 0.4	94.2 ± 0.4
autoencoder	73.0 ± 6.5	75.1 ± 3.5	84.2 ± 3.3	90.3 ± 1.5	94.3 ± 0.2
generative	73.4 ± 7.1	79.3 ± 4.1	**90.3 ± 2.4**	**95.4 ± 0.2**	97.1 ± 0.3
	Pathology-Specific Self-supervised Tasks				
magnification	76.3 ± 4.0	76.6 ± 3.6	87.4 ± 2.3	92.5 ± 0.2	94.1 ± 0.4
JigMag	**80.6 ± 3.5**	**81.8 ± 5.3**	89.5 ± 5.4	92.4 ± 0.5	96.5 ± 0.2
hematoxylin	75.3 ± 7.6	80.2 ± 5.3	87.5 ± 1.2	94.4 ± 1.3	**97.4 ± 0.5**
Best self-supervised	**80.6 ± 3.5**	**81.8 ± 5.3**	**90.3 ± 2.4**	**95.4 ± 0.2**	**97.4 ± 0.5**

Table 4.3: Camelyon16 Results for Different Annotation Budgets. Annotation budget is defined as the percentage of available WSIs that are labeled. The number of patches associated with each budget are indicated in the parentheses. The supervised upper bound performance when using all labeled data is 94.2%.

Labeled WSIs (No. Patches)	1%(600)	2%(1000)	5%(2600)	10%(6400)	20%(13540)
	AUROC(%)	AUROC(%)	AUROC(%)	AUROC(%)	AUROC(%)
	Baselines				
supervised baseline	68.3 ± 5.1	74.5 ± 5.8	81.2 ± 2.5	88.4 ± 2.3	92.1 ± 0.5
Mean Teacher [135]	73.7 ± 3.8	78.5 ± 2.6	84.5 ± 2.4	92.7 ± 1.9	93.1 ± 0.9
VAT [136]	70.9 ± 5.8	77.4 ± 3.3	81.3 ± 5.2	90.3 ± 2.3	92.8 ± 1.5
TS chain [150]	74.9 ± 6.9	76.9 ± 3.2	83.8 ± 2.1	93.1 ±2.5	**93.9 ± 1.3**
	Pathology-Agnostic Self-supervised Tasks				
rotation	69.8 ± 4.8	74.5 ± 3.1	80.4 ± 2.5	90.1 ± 2.0	92.4 ± 2.5
flipping	70.2 ± 6.2	75.4 ± 3.5	81.6 ± 5.1	89.4 ± 0.6	92.3 ± 1.6
autoencoder	70.1 ± 2.4	75.6 ± 4.1	82.3 ± 4.5	90.5 ± 2.3	92.4 ± 1.1
generative	72.5 ± 5.5	77.6 ± 5.4	82.4 ± 7.2	92.6 ± 3.2	93.6 ± 1.5
	Pathology-Specific Self-Supervised Tasks				
magnification	77.5 ± 3.1	84.6 ± 5.2	85.1 ± 3.6	**93.2 ± 3.4**	93.4 ± 2.5
JigMag	**81.7 ± 3.8**	**86.3 ± 5.2**	**87.4 ± 4.5**	90.6 ± 4.6	92.8 ± 2.4
hematoxylin	72.8 ± 4.6	78.3 ± 4.5	84.6 ± 3.4	92.3 ± 4.1	**93.7 ± 2.5**
Best Self-supervised	**81.7 ± 3.8**	**86.3 ± 5.2**	**87.4 ± 4.5**	**93.2 ± 3.4**	**93.7 ± 2.5**

Table 4.4: Kather Results for Different Annotation Budgets. Annotation budget is defined as the percentage of available WSIs that are labeled. The number of patches associated with each budget are indicated in the parentheses. The supervised upper bound performance when using all labeled data is 99.4%.

Labeled WSIs (No. Patches)	0.1%(100)	1%(800)
	AUROC(%)	AUROC(%)
	Baselines	
supervised baseline	87.5 ± 2.0	92.5 ± 1.2
mean teacher [135]	89.1 ± 1.5	93.9 ± 0.3
VAT [136]	88.5 ± 1.4	92.6 ± 0.4
TS chain [150]	88.9 ± 0.3	93.5 ± 0.2
	Self-supervised tasks	
generative	88.4 ± 3.5	92.3 ± 2.6
rotation	87.4 ± 1.6	93.3 ± 0.4
flipping	88.6 ± 0.8	93.0 ± 0.9
autoencoder	89.3 ± 1.3	94.3 ± 1.2
hematoxylin	90.3 ± 0.7	95.1 ± 0.5
Best self-supervised	**90.3 ± 0.7**	**95.1 ± 0.5**

achieve highest AUROC for any annotation budget, but it's performance is competitive with mean teacher and VAT. Similar to LNM-OSCC, JigMag could achieve highest performance overall, and the main boost is obtained at very low annotation budgets.

4.4.4.3 Results for Kather Dataset

We report performance of each of the self-supervised tasks on Kather dataset in Table 4.4. Since there are 9 classes in the Kather dataset, Macro AUROC is used for evaluation of classification performance. Unlike the other 2 datasets, only patches were available for this dataset, therefore the annotation budget only reflects the proportion of the overall patches that is labeled. Further, we observe that at 2% annotation budget, the performance of supervised baseline is still high (Macro AUC of 98%). Hence using semi-supervised approaches would not add much benefit. Hence, we focus on the very low annotation budget regime where some degradation of Macro-AUC can be observed for supervised model – i.e., annotation budgets of 0.1%(100 labeled) and at 1% (800 labeled images). Moreover, as this dataset does not include WSIs, we were unable to extract large patches or patches at different magnificationsand hence could not evaluate JigMag and magnification self-supervised tasks on this dataset.

From Table 4.4, we observe that at 0.1% annotation budget, predicting hematoxylin channel as a self-supervised task improves the performance by 2.8% and 1.2% compared to the baseline and mean teacher, respectively. At 1% annotation budget, we see that the various self-supervised tasks can again

Table 4.5: AUROC results for domain adaptation

	Cam16→LNM-OSCC	LNM-OSCC→Cam16
	Baselines	
supervised baseline	79.53 ± 0.2	63.73 ± 0.5
DANN	89.23 ± 1.5	71.15 ± 0.6
WDGRL	89.64 ± 2.6	72.65 ± 2.2
	Pathology-Agnostic Self-supervised Tasks	
rotation	86.14 ± 3.4	66.91 ± 4.1
flipping	82.14 ± 3.6	65.95 ± 4.4
autoencoder	89.90 ± 2.8	71.62 ± 2.6
generative	91.54 ± 3.5	74.14 ± 2.7
	Pathology-Specific Self-supervised Tasks	
magnification	89.69 ± 3.6	73.62 ± 4.1
JigMag	92.34 ± 4.4	74.51 ± 3.6
hematoxylin	90.47 ± 4.5	73.24 ± 3.8
mag+hem+JigMag	92.85 ± 3.6	74.95 ± 3.5

improve performance compared to the baseline. Predicting hematoxylin channel can also give the superior performance, suggesting that the prediction of rough nuclear segmentations can be helpful for semi-supervised learning.

4.4.5 Domain Adaptation Experiments

We conduct two domain transfer experiments, (i) Camelyon16 to LNM-OSCC (Cam16→LNM-OSCC) and (ii) LNM-OSCC to Camelyon16 (LNM-OSCC→Cam16). In both cases, we do unsupervised domain transfer, where the source is the labeled set and the target set is completely unlabeled.

We evaluate our approach against the naive supervised baseline, and two other domain adaptation methods WDGRL [154] and DANN [153]. The supervised baseline employs Resnet50 and is trained with source domain data only. WDGRL trains a domain critic network to estimate the Wasserstein distance between the source and target feature representations. The feature extractor network will then be optimized to minimize the estimated Wasserstein distance in an adversarial manner. By iterative adversarial training, WDGRL learns feature representations invariant to the covariate shift between domains. DANN is a domain prediction approach based on the GRL unit and was mentioned in Section 4.3.4.

We report the results obtained with Self-Path (using different pretext tasks) and the comparisons with the supervised and domain adaptation baselines in Table 4.5. We observe that the pathology-specific pretext tasks can help the model outperform the baseline by a large margin. For Cam16→LNM-OSCC, the pathology-specific pretext tasks provide more than 10% boost in AUROC over the supervised baseline. The combination of all pathology specific pretext

tasks achieves the best performance. Amongst the individual pretext tasks, JigMag achieves the best performance (∼2% better than DANN and WDGRL). Further, we note that the pathology agnostic generative model also performs well – with 1.9% higher AUROC than WDGRL and 11% higher AUROC over the supervised baseline. This suggests that the images from the generator can contribute to learning useful domain-invariant features as well. We see similar trends for LNM-OSCC→Cam16 – where again combining pathology specific tasks has the best performance and JigMag provides the second best performance. We highlight that we have used domain prediction with GRL layer in all non-generative methods as it improves the performance. Generative models, owing to adversarial training can still achieve very high performance, even without GRL. The fully supervised non-domain adapted AUROC values for LNM-OSCC and Cam16 are 98.2 and 95.3, respectively.

4.4.5.1 WSI Analysis

While the results thus far are reported at the patch level, it is also useful to consider the WSI-level performance. For the Cam16→LNM-OSCC domain adaptation task, we now report the WSI-level results for the top two best performing Self-Path settings i.e., combination of all pathology specific pretext tasks and JigMag pretext task. We also provide comparisons with the supervised baseline (source only), WDGRL, and the pathology agnostic generative pretext task.

In order to quantify WSI-level performance, we aggregate patches belonging to a WSI and construct a WSI-level heat map based on the patch level predictions. For heat map generation, there are two steps. First, we extract patches of 128 × 128 at 10× magnification with overlap of 50% from tissue regions of WSIs. Second, we aggregate the prediction of each patch together to build the final heat map of WSIs. We then post-process these heat maps to obtain the WSI-level prediction. The post-processing steps are uniform for all models in this section, and as follows: we extract 10 morphological and geometrical features from objects within binarized heat map at three thresholds of 0.25, 0.5 and 0.9. Then we calculate the mean, stddev, minimum and maximum of object features for each WSI. Therefore, in total we use 120 features for constructing feature vectors. Afterwards, we employ the random forest algorithm for classification of the features. Finally, we evaluate the model on the test set of LNM-OSCC.

The results are shown in Table 4.6. The supervised baseline has WSI-level AUROC of 75.2% whereas Self-Path with JigMag pretext task and Self-Path with the combination of all pathology specific pretext tasks each improve the performance by 16.4%. Further, we note that Self-Path with JigMag improves

Table 4.6: Cam16 → LNM-OSCC domain adaptation results on the WSI-level. The upper bound performance using all labels for target domain in supervised fashion is 93.3%.

	AUROC(%)	Average Precision(%)
supervised baseline (source only)	75.2	81.7
WDGRL	85.8	91.6
generative	90.4	95.2
JigMag	91.6	96.7
mag+JigMag+hem	91.6	96.3

performance over WDGRL by 2% at the patch-level and a ∼6% improvement at the WSI-level. This suggests that the magnification puzzle and the pretext tasks that can help learn from various image resolutions in a self-supervised manner enable strong performance boost at WSI-level (beyond patch-level).

These improvements are also evident in the WSIs overlaid with the heatmaps, as visualized in Figure 4.4. This figure shows that the supervised baseline (source only) model (middle column) has many false negatives and often misses tumor regions. However, WDGRL, Self-Path with JigMag, and Self-Path with generative pretext task can all increase true positives while decreasing false negatives. We note that WDGRL and Self-Path with generative pretext task do not perform as well as Self-Path with JigMag - mainly because they suffer larger number of false positives at the patch-level classification.

4.5 Discussion

In this section we describe sensitivity analyses and discuss the model performance by changing the values of loss weights, decreasing the annotation budget and combing all pathology specific tasks. Moreover, we conduct an experiment to show the usefulness of transfer learning using our proposed self-supervised tasks.

4.5.1 Effect of Loss Weight for Each Task

We consider the task of training with 1% of annotation budget on Camelyon16 dataset. To understand the effect of loss weights for each pretext task, we experiment with different values of α and show the results in Table 4.7. Overall, assigning more weights on each task shows better performance. More precisely, when α is set to 1, maximum value of AUROC is obtained. Therefore we can conclude when we are using only one pretext task, the pretext task and the main task should have similar weight to be effective for semi-supervised learning. The optimum value of α may change when we use all tasks together which we investigate in the next section.

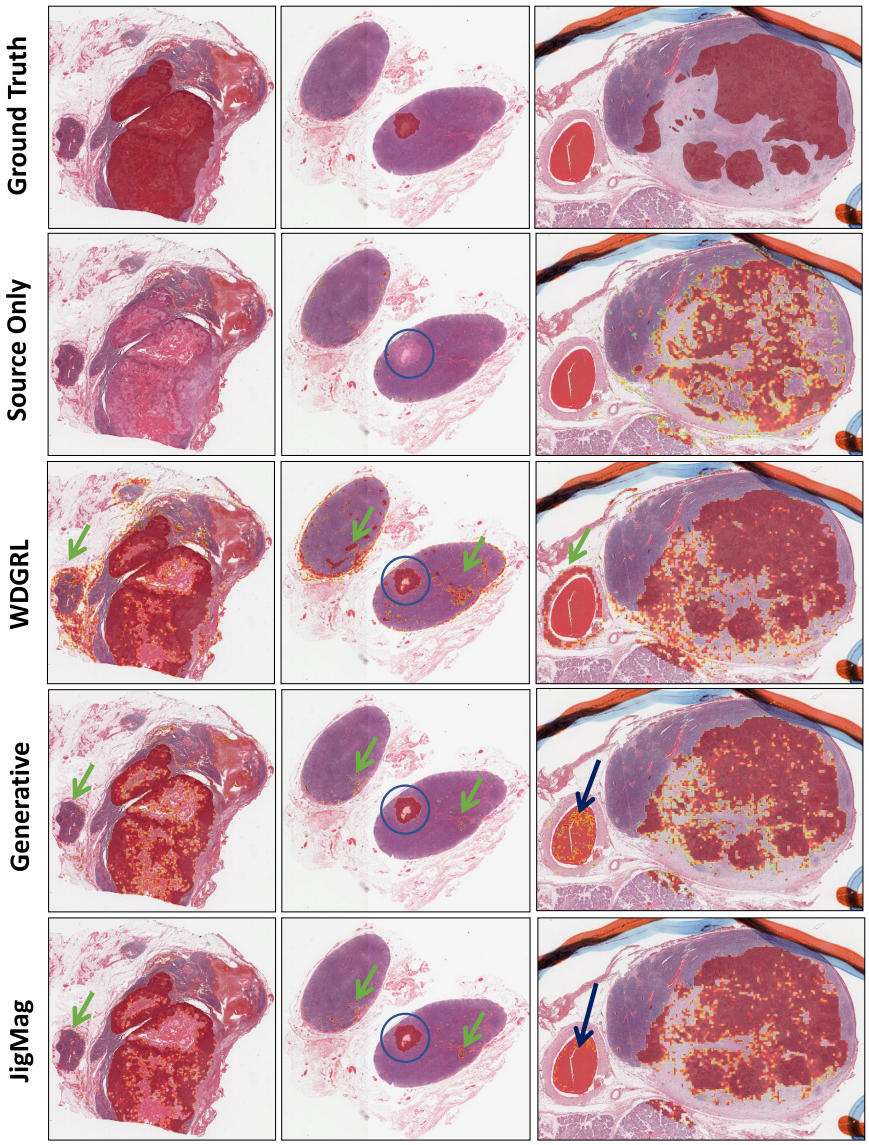

Figure 4.4: Three WSI samples and their overlaid heatmaps. from top to bottom, first row: the overlaid ground-truth mask, second row: overlaid heat map of model predictions when it is trained using only Camlelyon16 data, third row: Overlaid heatmap of WDGRL predictions, fourth row depicts the overlaid predictions of Self-path using generative task and the last row shows the heatmaps generated Self-path using JigMag task. the The circle indicates a region which is missed using the supervised baseline (source only) model and green arrows point to the false positive regions generated by WDGRL where using generative task and JigMag task eliminate those regions. Black arrow also shows regions that are misclassified by generative model but are correctly classified as normal regions by Jig-Mag. (Best viewed in color, zoom in to see more details)

Table 4.7: AUROC performance of pathology specific tasks with different values of α on Camelyon16 dataset.

α	magnification	JigMag	hematoxylin
1	77.5 ± 3.1	81.7 ± 3.8	72.8 ± 4.6
0.8	77.1 ± 2.8	81.5 ± 3.4	71.3 ± 2.4
0.6	76.4 ± 4.0	78.8 ± 2.6	70.2 ± 3.5
0.5	74.6 ± 3.4	78.4 ± 2.4	70.3 ± 4.6
0.2	72.5 ± 3.7	74.1 ± 4.6	69.5 ± 4.4

Table 4.8: Using all pathology specific tasks for semi-supervised learning on Camelyon16 dataset. α_{mag}, α_{JigMag} and α_{hem} indicate the loss coefficient for magnification, JigMag and hematoxylin tasks, respectively.

α_{mag}	α_{JigMag}	α_{hem}	1%	2%
1	1	1	79.1 ± 4.5	83.5 ± 5.1
0.25	0.5	0.25	83.2 ± 4.3	86.3 ± 5.3
0.5	0.25	0.25	80.2 ± 2.5	85.4 ± 3.1
0.25	0.25	0.5	79.6 ± 2.7	84.3 ± 5.5
0.25	0.25	0.25	80.3 ± 3.4	85.5 ± 1.8

4.5.2 Combining tasks

We now evaluate the effect of the loss weights (α's) when combining all pathology specific tasks. We consider the task of training with 1% and 2% of annotation budget on Camelyon16 dataset, and experiment with different combinations of loss coefficients. The results, in Table 4.8, suggest that assigning high weights (similar to main task) to all pretext tasks can degrade the performance. For example, if all tasks are given $\alpha = 1$, overall the weights for pretext tasks would be 3× more than the main task which would cause drop in performance. However by assigning smaller weight values for each task, we can achieve better performance. Particularly, best performance is obtained when more weight is assigned to JigMag task and lower weights to Hematoxylin and magnification tasks. This is in line with previous experiments which showed that JigMag had better performance as compared to other tasks. We can, therefore, recommend that a good strategy can be to start with heavy weight to JigMag for computational pathology tasks before combining it with other self-supervision tasks.

4.5.3 Performance at Very Low Annotation Budget

In Section 4.4.4, we evaluated the performance of self-supervised tasks with different annotation budgets. we observed, despite high boost in performance by applying self-supervised tasks, the supervised baseline also gives reasonable results (e.g., 73.4% on LNM-OSCC for 134 patches). To assess performance at even lower annotation budget, we further decreased number of patches

Table 4.9: AUROC results for very low budget of annotation:here only 25 image patches are used in each class.

	Camelyon16	LNM-OSCC
	Baselines	
supervised baseline	55.3 ± 5.1	54.8 ± 8.1
mean Teacher	65.4 ± 4.8	60.4 ± 5.4
VAT	64.3 ± 6.4	58.6 ± 6.5
TS chain	62.4 ± 10.6	59.4 ± 7.7
	Pathology-Agnostic Self-supervised Tasks	
rotation	62.6 ± 4.6	58.7 ± 4.6
flipping	65.7 ± 9.3	58.9 ± 5.3
autoencoder	65.1 ± 6.4	59.6 ± 4.3
generative	64.2 ± 5.7	60.1 ± 10.3
	Pathology-Specific Self-supervised Tasks	
magnification	65.3 ± 7.5	62.2 ± 6.7
JigMag	**66.2 ± 6.4**	**63.5 ± 7.9**
hematoxylin	64.2 ± 7.4	62.4 ± 4.6
mag+hem+JigMag	**66.5 ± 5.5**	**64.1 ± 5.5**

annotated (while maintaining the same number of WSIs) to 50 for LNM-OSCC and Camelyon datasets. As shown in Table 4.9, Self-Path with pathology-specific pretext tasks can improve the AUC by about 10% over the supervised baseline. Again, the JigMag pretext task is the best performing pretext task. Moreover, we also note that combining all pathology specific tasks (with loss weights 0.25, 0.25 and 0.5 for hematoxylin, magnification and JigMag respectively) can result in even better performance.

4.5.4 Transfer Learning

We finally investigate the usefulness of the representations learned by Self-Path for related tasks. For this, we conduct a transfer learning experiment using Camelyon16 dataset. We first train Self-Path with each self-supervised pretext task on the entire dataset, and then fine-tune the backbone (the model excluding the final linear layer/decoder) for the main task. We compare the performance against the naive method of training the network from scratch with random weight initializations (Scratch). The results for different pretext tasks at varying annotation budgets are shown in Table 4.10. We can see that the representations learned by Self-Path with transfer learning enable performance improvement over 'Scratch' in each case. Again, Self-Path with JigMag achieves the best performance. The improvements with fine-tuning is largest in the low annotation regime, and drops off when more annotated data are available. These results suggest that the pretext tasks in Self-Path enable learning of useful representations. Overall, with annotation budget of over 20%, fine-tuning gives the same result as training from scratch. This

Table 4.10: Results of transfer learning of self-supervised tasks with different budget of annotations using Camelyon16 dataset.

	1%	2%	5%	10%	20%
Scratch	68.3	74.5	81.2	88.4	92.1
magnification	72.6	77.4	84.8	89.9	92.2
JigMag	73.3	79.4	85.8	90.4	92.7
hematoxylin	72.9	79.5	85.9	88.6	92.3

phenomenon is also shown by [172].

4.6 Summary

In this chapter, we proposed Self-Path – a generic framework based on self-supervision tasks for histopathology image classification – to address the challenge of limited annotations in the area of computational pathology. We introduced 3 novel self-supervision tasks to cater to the contextual, multi-resolution and semantic features in pathology images. We showed that such pathology specific self-supervision tasks can improve the classification performance for both semi-supervised learning and domain adaptation. Moreover, we thoroughly investigated general self-supervised approaches such as generative models within this pipeline and showed that using the pathology-specific tasks, despite being simple and easy to implement, can improve performance over generic self-supervision in many scenarios involving limited annotation budget or domain shift. In particular, we note that the JigMag self-supervision can be extremely helpful when the amount of labeled data is very small. Unlike baseline methods that are highly dependent on hyperparameters values, our method can achieve good performance without exhaustive hyperparameter tuning. Self-Path can be applied to other problems in computational pathology, where annotation budget is often limited or large amounts of unlabeled image data are available. Other future directions include employing other self-supervision tasks (such as predicting the Eosin channel or a combination of Hematoxylin and Eosin after estimating the two channels, rather than keeping them fixed), increasing the number of magnification levels, adding the magnification as extra augmentation to supervised baseline, increasing the JigMag grids to incorporate wider and more complex puzzles for the network to solve and a deeper investigation into other domain adaptation tasks. It is also worth mentioning that most of semi-supervised learning approaches (specially the ones proposed in this paper) are useful when the budget of annotation is very low. They can not contribute when we have large annotation budgets. Therefore designing methods that can push the margin even while having enough annotation can be an interesting future direction. For detailed discussion of

current semi-supervised learning approaches interested readers are referred to [173].

Chapter 5

Predicting Non-Small Cell Lung Cancer Survival

5.1 Introduction

Lung cancer is the most common cancer related mortality worldwide. Non-small-cell lung cancer (NSCLC) accounts of 80% of all lung cancer types. Two major NSCLC types are Adenocarcinoma (ADC) (40%) and Squamous cell Carcinoma (SCC) (25-30%).

The five-year survival rate of lung cancer is 17.7%, which is lower than that of many other leading cancers, such as colon cancer (64.4%) and breast cancer (89.7%) [174]. Accurate survival analysis is necessary for personalized treatment management and prognosis. Therefore, predicting clinical outcome of lung cancer is an active field in today's cancer research. Histopathology images serve as the gold standard for diagnosis of lung cancer and are primarily evaluated by pathologists or doctor. Most of current pathology diagnosis is still based on subjective opinions of pathologists and the varying abilities of doctors could result in large interpretation errors or bias. Pathologists make diagnostic decisions based on cellular and inter-cellular level morphology, and thus accurate cell localization/segmentation is a prerequisite step for lung cancer survival analysis [174]. It has been shown that there exists connection between lung tumour morphology and prognosis [175]. Emergence of Whole slide Images (WSIs) have brought many opportunities and challenges for analysis of tumour micro-environment. One main opportunity is allowing the Computer Aided Diagnosis (CAD) systems to be applied on them for fast and more precise diagnosis. One main challenge is the large size of these images where one needs to chunk them into small image patches to be able to process them by machine learning models.

Wang et. al [176] extracted 3 groups of geometry, texture and pixel intensity statistics features from images and then the Cox proportional hazards model

was used to select 166 image features that are correlated with patient survival outcome . Yao et al. [177] extracted features from three cell subtypes (tumor, lymphocyte, stromal) and developed a survival model for two subtypes of NSCLC: ADC and SCC. Yu et al. [178] showed that textural and morphological features extracted from tumor nuclei and tumour cytoplasm of each image patch can predict survival for bot ADC and SCC. They used CellProfiler [179] for segmatiation and extracting features. In [180], radimoic and pathomic features are combining for predicting recurrence in early stage lung cancer. Cheng et. al [181] showed that there is a correlation between patient survival and topological features.They have used a deep auto-encoder to cluster cell patches into different types and then construct a graph for each patch to extract topological features.

Deep learning based-approaches learn feature representations in an end to end manner. Early neural network models have been applied to the problem of survival analysis which models the nonlinear survival data [182–186]. These models have not outperformed standard methods for survival analysis, since the neural networks were not developed as they are today. Yousefi et. al [187] showed for the first time that performance on high-dimensional data of a Cox neural network can have competitive performance for survival analysis. Katzman et al. [188] showed that Cox proportional hazards deep neural network can have state-of-the-art performance on low-dimensional data. Inspired by [188], various methods have been developed in computational pathology to predict risk for histology image or genetic data.

WSIS approach [189] extracts hundreds of patches from each WSI by adaptive sampling and then group these images into different clusters. Then an aggregation model is trained to make patient-level predictions based on cluster-level Deep Convolutional Survival (DeepConvSurv) output, Yao et. al [190] combined genome modality with DeepConvSurv for survival prediction using multi-modality data Yao et. al [191, 192] use multiple instance learning to encode all patches and different patterns in the WSIs to predict survival where they have shown their model to be capable of predicting risk score for two datasets of lung and brain. Multiple instance learning is a proper choice for predicting survival based on the WSIs because one label is only provided for a WSI or a collection of WSIs belonging to a patient. Therefore by presenting the WSIs as a bag of instances (image patches/features), this problem can be solved by using multiple instance learning approaches. This is discussed in detail in the following sections.

Most approaches rely on exhaustive annotations of ROIs to extract features or to train deep models which is not often possible. The success of these approaches mainly depends on the selection of representative patches and integration of patch prediction to come up with the final output. Manual

annotations of regions of interest and/or selecting the representative patches of WSIs is challenging and is not an optimal approach. Extensive and time-consuming manual annotations in clinical practice is an uphill task. Moreover, most deep learning based approaches are based on deep features which are hard to interpret. Therefore, to address the deficiencies of current models, recent approaches for analysis of WSIs rely on the success of weakly supervised learning particularly multiple instance learning [193]. In this chapter by concentrating on the explainability and reducing the need for manual annotation, we propose attention based multiple instance learning to extract the representative patches from WSIs and then investigate the morphological features of these representative patches for their potential association with patient survival.

In the following, we will first cover basics of survival analysis in Section 5.2; afterwards, the details of our framework including multiple instance learning, segmentation, Cox model are explained in Section 5.4. Finally, we discuss the results in Section 5.7.

5.2 Survival Data

Survival analysis is concerned with predicting the time until an event occurs, such as onset of a disease, tumor recurrence, death after some treatment intervention, etc. Survival data has three components: 1) *Time* T from the beginning of follow-up of an individual until an event occurs, 2) *Event* E which is a designated experience of interest that may happen to an individual (death, recurrence, relapse, etc.), and 3) Patient *data* x whose association we are trying to explore with patient survival. In survival data, all observations do not always start at zero; in other words, we do no need exact starting points and end points. Starting point of the study determine the starting point of survival time for all subject (all the durations are relative,). If the event (e.g. death) is observed, the time interval T is associated with the elapsed time between the starting point of study (the time in which the data was collected) and the time that event occurs, and the event indicator is $E = 1$. If the event is not observed the time interval T corresponds to the elapsed time between the start of study and the last contact with the patient (end of study, patient died due to reason the are not related to study, patient withdraw from study, etc.), and the event indicator is $E = 0$.

There are different types of censoring, but the most common one is *right censoring* where we only have the information about the patient up to a certain time and the information after that time is unknown.

Event or censorship is an important issue which needs to be taken into account for modeling survival data. Moreover, probability of survival for each patient decreases as we move toward the end of study. Therefore, standard

statistical methods such as linear regression are not suitable for survival time data. Cox proportional hazard model (CPH) [194] is the most commonly used method in survival analysis.

Survival function and hazard are two fundamental functions in survival analysis. Survival function $S(t) = \Pr(T > t)$, gives the probability that a person survives longer than some specified time t. Hazard function is a measure of risk at time t, which gives the instantaneous potential per unit for the event to occur, given that individual has survived up to time. The hazard function focuses on the failing. A proportional hazards model is a common method for modeling an individual's survival given their covariates (x). This model gives an expression for the hazard at time t for an individual with a given specification of a a set of explanatory variables denoted by X. X represents a collection of predictor variables that is being modeled to predict an individual's hazard.

$$h(t, X) = h_0(t)e^{R(x)}$$

where $R(x)$ is risk function denoting the effects of an individual's covariates and h_0 is the base hazard function depending on the time. For *linear survival models*, the CPH is a proportional hazard model that estimates the risk function $R(x)$ by a linear function $\hat{R}(x) = \sum_{i=1}^{p} \beta_i x_i$. And the goal is to find the weights β to optimize the Cox partial likelihood. The partial likelihood is the product of the probability at each event time T_i that the event has occurred to individual i, given survival up to this time. The Cox partial likelihood is defined as:

$$L_c(\beta) = \prod_{i:E_i=1} \frac{\exp(\hat{R}(x_i))}{\sum_{j \in \Re(T_i)} \exp(\hat{R}(x_j))} \tag{5.1}$$

where T_i, x_i and E_i represent the event time, covariates and the event indicator for the ith observation. The risk set $\Re(T_i) = \{i : T_i \geq t\}$ is the set of patients still at risk of failure at time t and the product is defined over the set of patients with $E_i = 1$.

In *non-linear* survival models, the risk function $\hat{R}(t)$ is the output of neural network and the input is either covariates or the raw input (image). The loss function for the model is defined as the negative log likelihood of Eq. (5.1):

$$l(\theta) := -\sum_{i:E_i=1} \left(\hat{R}(x_i) - \log \sum_{j \in \Re(T_i)} e^{\hat{R}(x_j)} \right) \tag{5.2}$$

In this chapter, we consider both linear and non-linear models and we observe that linear models using morphological features as survival covariates are prognostically important. Based on our experiments on NSCLC dataset from TCGA repository, deep neural network in a weakly supervised manner

Table 5.1: Clinical features for Lung cohort of TCGA and log rank test p-value for disease specific survival (DSS).

Categorical Clinical Features		Count	Percentage	DSS p-value
Number of patients		778	100%	-
Gender	Female	309	40%	0.83
	Male	469	60%	
TNM Stage	Stage 1	419	52%	3e-9
	Stage 2	211	28%	0.02
	Stage 3	111	16%	1e04
	Stage 4	25	3%	6e-5
	Not Reported	12	1%	-
Cancer Type	LUAD	396	50%	0.08
	LUSC	382	50%	
Patient Status	Alive	518	60%	-
	Dead	260	40%	
Continuous Clinical Features	Mean	STDDEV	Median	
Age (year)	66.54	9.45	68	0.8
Survival (month)	31.77	31.80	21.76	-

with negative log of Cox partial likelihood do not perform well. In the next section we describe the dataset details.

5.3 Dataset

We have obtained our lung dataset from The Cancer Genome Atlas (TCGA) repository. There are 1054 lung diagnostic WSIs belonging to 924 patients in this repository. We have removed images with artefacts such as pen marking, blured and folding. Images with very small tissue regions and lost associated disease specific survival time or event indicator are also not considered in our study. Overall, we are left with 778 cases and 824 WSIs. The disease-specific survival (DSS) time is the elapsed time between the beginning of an individual follow up and an individual death or the last follow-up in case of censored data. In TCGA cohort, most of the patients were diagnosed between 1992 and 2013, and the average age and median age of patients are 66.54 years and 68 years, respectively, with standard deviation of 9.45. The minimum and maximum age in this cohort are 33 and 88 years. Number of male patient (469) is more than female patients (309). The distribution of TNM-stage of cases is slightly skewed toward lower stage with 52% cases of stage I. The two NSCLC subtypes are alsmost equally distributed. 40% of patients died during the study where 23% of the number of deaths are disease specific ($E = 1$). The average and median of DSS time of all patients is 31.77 and 21.76 months, respectively. The details of patients statistics are reported in Table 5.1.

5.4 Method

In our framework we encode the extracted patches from WSIs using an off-the-shelf deep network, then encoded images (features) for each WSI form a bag which is processed by an attention-based Multiple Instance Learning (MIL) approach. MIL is trained to classify the bags to distinguish between cancer subtype of WSIs. After the model learned the underlying representation of data, attention weights are used to select the top k patches to extract the morphological features. Afterwards, the morphological features are used in a Cox proportional hazard (CPH) model to predict the risk score. The risk value obtained from our pipeline is called representative patch morphology (RPM) score and we show that RPM score is prognosticator of clinical outcome of NSCLC. We go through the details of our pipeline for predicting outcome for patients with NSCLC. Overview of our pipeline is shown in Fig. 5.1.

5.4.1 Image Patch Encoding

Each WSI contains thousands of small 224×224 patches, where considering all these information at once for processing is cumbersome and will need high memory and computational resources. Random extraction of patches is not a good solution as we might lose some important information. Therefore, we extract all patches from WSIs at $20\times$ magnifiction ($0.5\mu m$ per pixel), then we encode all images patches belonging to tissue regions by using ResNet18 [56] to reduce the input dimension. Image ResNet18 is pre-trained on Imagnet and the features are taken from last convolutional layer after applying global average pooling. Each 224×224 patch is transformed to a vector of length 512. This amount of dimension reduction enables us to represent all tissue region patches of WSI to the model as is described in the next section.

5.4.2 Attention-based Multiple Instance Learning

MIL is a weakly supervised learning approach where a single label is assigned to the bag of instances [195]. And the label of the bag is positive ($Y = 1$) if at least one instance in the bag is positive. Since exhaustive annotations of WSIs is not possible at large scale, MIL has been recently utilized in computational pathology to predict a label for a WSI, where a WSI is considered as a bag of instances (e.g. bag of image patches or bag of features). In MIL, one important challenge is finding key instances. Key instances are the instances that contribute more to predict the label for a bag. Key instances are of high importance as they can help to interpret the final decision and may give more insight about the underlying relation between a diagnosis and tumour micro-environment.

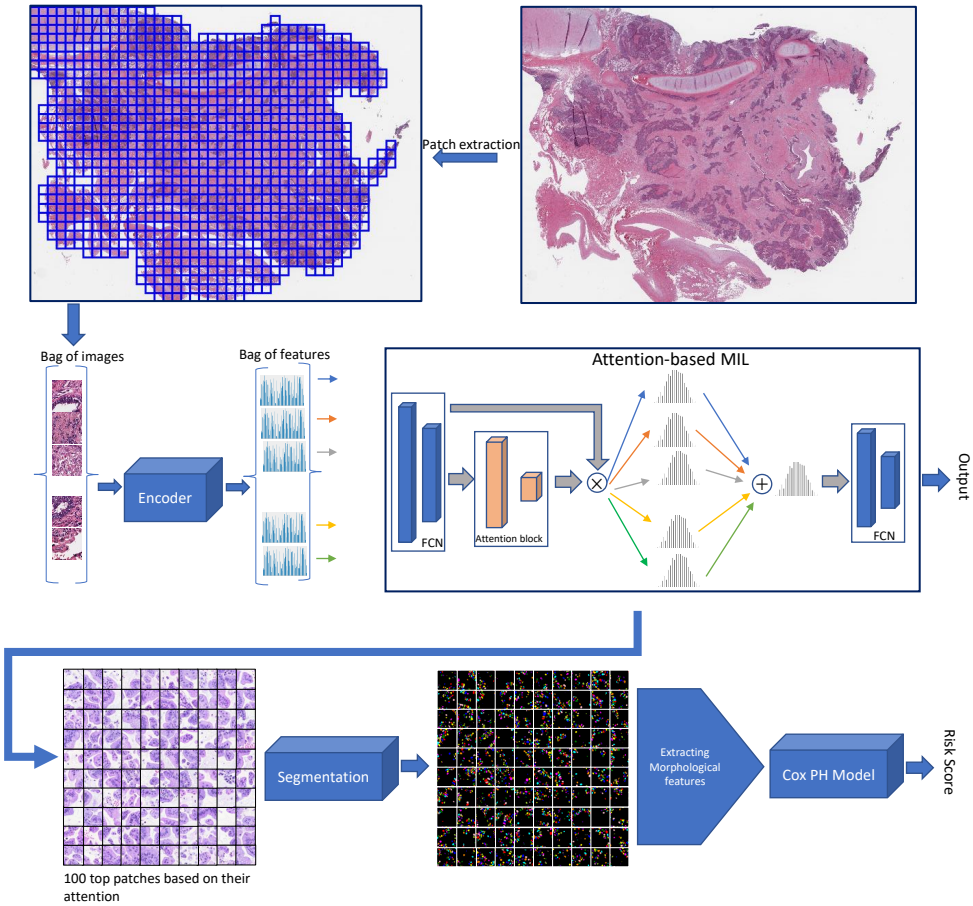

Figure 5.1: Schematic overview of our framework.

One approach to MIL is training a classifier on instances where each instance has the label of its corresponding bag, and then aggregate the scores by permutation-invariant operators such as maximum, mean, etc. Another approach is to map instances to a low-dimensional embedding and then classify them by a bag-level classifier. The latter approach is preferred because it learns the joint representations of bag instances and introduce less bias compared to instance-level approach where individual instance does not have precise label. Moreover, in the context of neural networks, embedding-based approaches can be trained end to end.

Aggregation of output scores for each instance or aggregation of embedding is referred to as pooling. Attention based MIL pooling is of high interest because of being trainable and the interpretability it offers. More precisely, let $H = \{h_1, h_2, ..., h_k\}$ be the embedding of bag of instances $X = \{x_1, x_2, ..., x_k\}$. The embedding set is obtained by a neural network, and the MIL pooling is a weighted average of embeddings where weights are determined by a neural network:

$$\begin{aligned} z &= \sum_{k=1}^{K} \alpha_k h_k \\ \alpha_k &= \frac{\exp(w^T \tanh(V h_k^T))}{\sum_{j=1}^{K} \exp(w^T \tanh(V h_j^T))} \end{aligned} \tag{5.3}$$

where $w \in {}^{L\times 1}$ and $V \in {}^{L\times M}$ are parameters. Hyperbolic $tanh(.)$ introduce non-linearity to the attention function. Since the weights are normalized, their value can show the contribution of each instance/embedding to the generation final output. Therefore, ideally high attention should be assigned to instance that are likely to be positive inside bag. In our pipeline attention plays an important role because it helps us to select the most representative patches in the WSI and therefore we can apply further analysis on those patches rather than on all the patches belonging to a WSI (these patches should be potentially tumour patches). Therefore, we build attention-based MIL where we try different labels and loss functions to train the model. Afterwards, the model with highest performance is used to select the representative patches. More precisely, we consider following experiments: 1) we train MIL with the negative log of Cox partial likelihood as loss function where the survival time and event values are used as labels, 2) we binarize the survival times by using their median value as threshold and cross entropy loss is used as loss function. for this experiment only observed events are considered for training ($E = 1$), 3) We consider the cancer stage as target variable where cross-entropy is used as objective function, and 4) we use the cancer subtype (ACC and ADC) as the label and train a MIL to distinguish between these two. Cross entropy is also used for this experiment.

MIL is used for binary classification where a bag is considered positive

when at least one sample in a bag is positive, and a bag is considered negative when all instances in a bag are negative. The definition of MIL may not hold for our classification and segmentation tasks where we are classifying cancer subtypes, because defining positive or negative instances are opaque. For such tasks the network looks for particular pattern in the bag which are indicative of the whole bag label. One may argue that our approach is aggregation learning or attention learning, however since we use the framework and the model that has been used for MIL, we stick to the MIL name.

In our experiments, training MIL to predict risk using Cox partial likelihood, binarized survival and prediction of stage did not give good performance. For risk prediction, we have used concordance index as evaluation metric, and Area Under the Receiver Operating Characteristic (AUROC) is used for the other three experiments. The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the true positive rate (TPR) against false positive rate (FPR) at various threshold values. In clinical studies, the concordance index gives the probability that a randomly selected patient who experienced an event (e.g. a disease or condition) had a higher risk score than a patient who had not experienced the event. It is equal to AUROC and ranges from 0.5 to 1. Fig. 5.2 shows the classification performance for binary survival prediction and cancer subtyping. As we could achieve the best performance for cancer-subtyping, therefore we conclude that the attention weights can be used for further analysis. We expect that the representative patches highlight the tumour regions and the specific characteristics of ACC and ADC that separate them from each other. In figure, we have shown two random WSIs and their attention weights overlaid on them. We can observe from this figure that patches with highest attention are mostly belong to tumour region and therefore we can use them for predicting survivals in the next steps. For 3 fold cross validation experiment, the MIL that we have used could achieve concordance index of 0.54 for predicting risk and AUC of 90%, 62%, 58% for cancer subtyping, binarized survival prediction and stage prediction ,respectively. As we later will see, cancer stage is a good predictor of patient outcome, therefore if we can predict patients stage, we can stratify them into poor and good prognosis cohorts. By predicting stage we did not achieve a satisfactory performance. The AUC for stage prediction is 55%.

5.5 Segmentation

After finding the most representative patches within WSIs, we segment the nuclei to extract morphological features. These morphological features serve as covariates for Cox model. In chapter 2 we have introduced SpaNet, a method for nuclear instance segmentation. Here we compare different nuclear

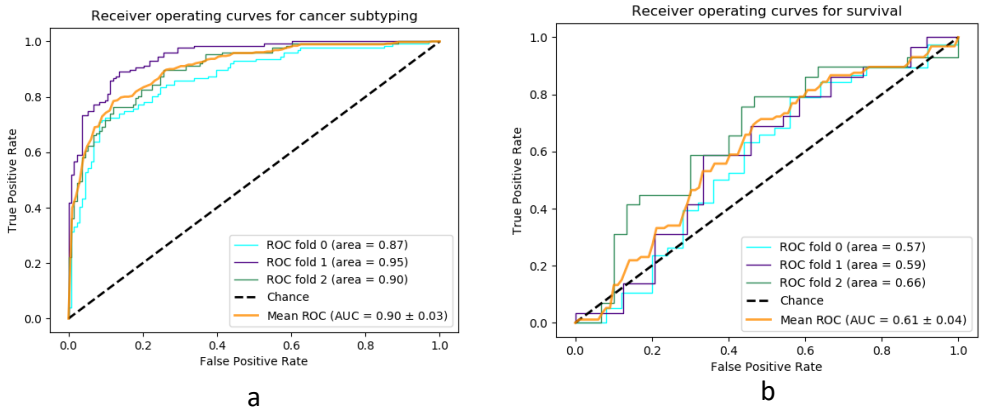

Figure 5.2: The classification performance of MIL models for a) cancer subtyping and b) survival prediction

Table 5.2: Comparative performance of different models on PanNuke dataset.

	Dice	Panoptic Quality
HoVerNet	0.8368	0.6596
Micronet	0.8028	0.6053
DIST	0.7523	0.5346
Mask-RCNN	0.6936	0.5528
SpaNet	0.8412	0.6604

segmentation methods on a broader dataset, PanNuke [196, 197] and then apply the best model on selected patches. PanNuke is a dataset of nuclear segmentation of 19 different tissue types where 7000 nuclei belong to the lung tissue. Therefor, it is a good choice for training segmentation models to achieve high generalizability. We compare SpaNet with HoVerNet [198], DIST [51], Mask-RCNN and MicroNet [199]. In Table 5.2, we have reported the performance of these models in terms of panoptic quality and Dice. As shown in the table, SpaNet and HoVerNet have the best performance where SpaNet is marginally better than HoVerNet with dice index of 0.8412 and PQ of 0.6604. Therefore, we have used SpaNet to segment the patches for further analysis. We have also shown in Fig. 5.4 some visual results of SpaNet on some random representative patches.

5.6 Morphological Features of Nuclei

We extracted overall 68 features for each case. All of these features are based on the results of nucleus segmentation. These features are extracted from each nucleus within a patch: area, area of bounding box, eccentricity, diameter of an equivalent circle that encompasses the nucleus, extent, length of major axis, length of minor axis, orientation of nucleus, perimeter and solidity. Statistics of Gray-Level Co-occurance Matrices (GLCMs) of each patch is also considered

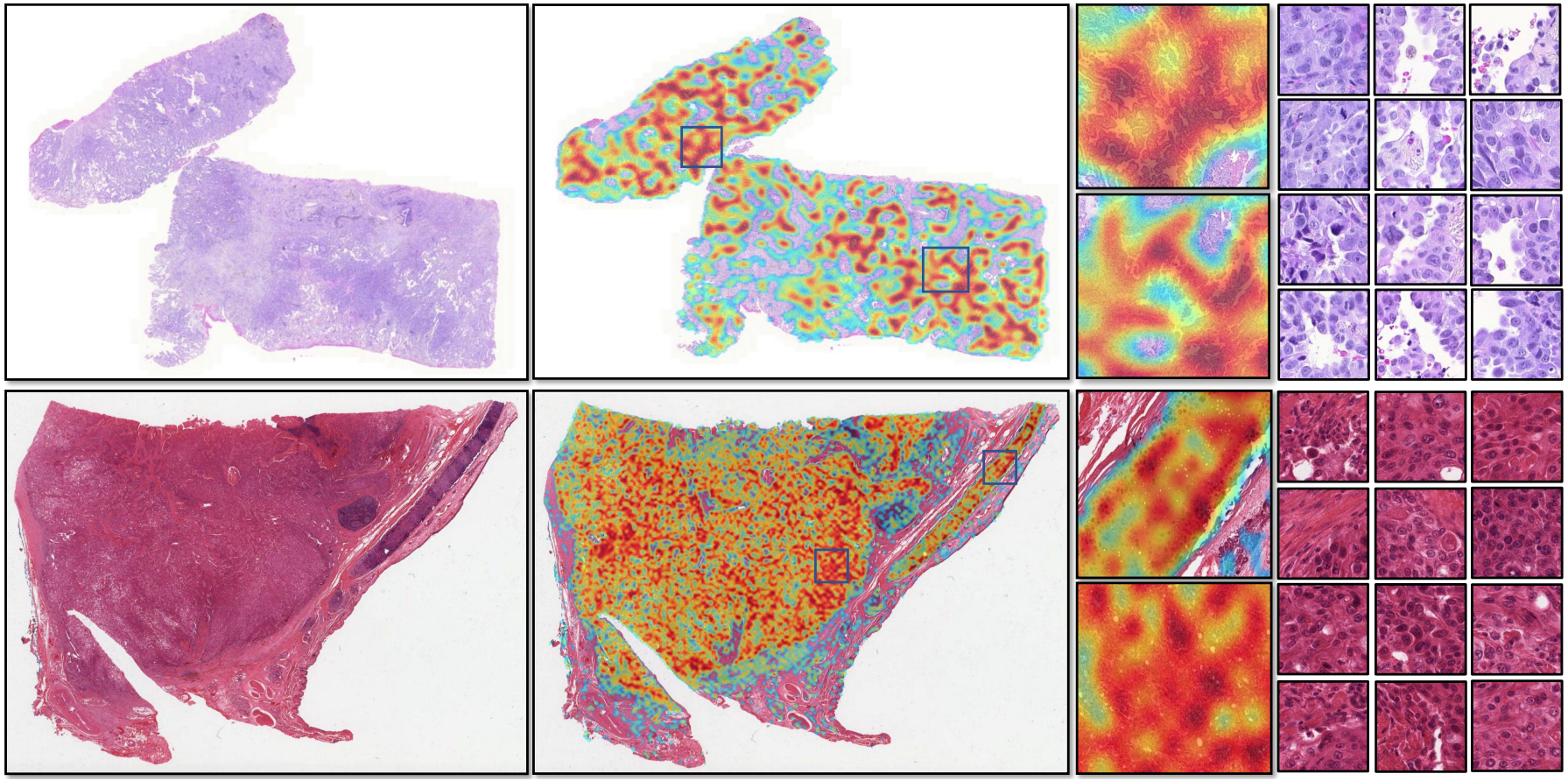

Figure 5.3: Two random selected WSIs and their corresponding attention map overlaid on them. Patches with high attention values mainly are from tumour regions.

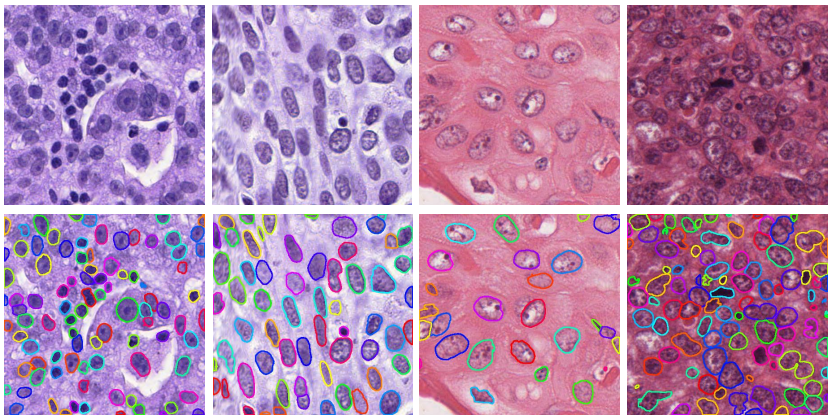

Figure 5.4: Visual segmentation output of Spa-net on 4 random representative patches.

in our feature vector. The GLCM calculates how often a pixel with gray-level value i accrues either horizontally, vertically, or diagonally to adjacent pixels with the value j. Here 4 directions are considered for calculating GLCM. These directions are horizontal (0), vertical (90), bottom left to top right (-45) and top left to bottom right (-135) where pixel offset is 1 in all case. we have derived contrast, correlation, energy, and homogeneity statistics from GLCM. Moreover, minimum and maximum intensities of each nucleus are also considered.

To construct the feature vector for each case, we compute the average, minimum, maximum and standard deviations of the values of nucleus based features for each patch. Consequently, the obtained values are averaged over 100 selected patches for each case, which overall sum up to 68 features. We standardize (scale between 0 and 1) the features before using them for Cox model. These features are then used in Cox proportional hazard model as described in Section 5.2. We find the coefficients (β) and the importance of each covariates based on the value of coefficients. Therefore, the risk score from Cox model is used as the final score for patient stratification and exploring its prognostic value.

5.7 Results and Discussion

In this section we are going to find the importance of features in predicting survival and investigate if the risk score predicted by model can stratify patients into two groups of high risk and low risk. To accomplish the first task, Cox's proportional hazard model is often a good choice, because its coefficients can

be interpreted in terms of hazard ratio, which often provides valuable insight. Moreover, the value of coefficients can be used to determine the importance of each feature. However, simple Cox model has a major drawback. When number of predictors increase, it might fail, because it internally tries to invert a matrix that become non-singular due to correlation among features. The penalized Cox models can address this issue. Ridge and Lasso penalty solve the mathematical problem of fitting a Cox model. The Lasso penalty is a good option if we want to select a subset of features that are predictive and ignore the remaining features.

Lasso is not a good choice for high dimensional data where number of features is more than number of samples, because it can not choose more features than number of sample. Moreover, if the features are correlated, the Lasso penalty randomly chooses one feature. The Elastic-net penalty solves this issues by combing weighted Lasso and Ridge penalty terms:

$$\arg\min_{\beta} \quad l(\theta) + \alpha \left(r \sum_{j=1}^{p} |\beta_j| + \frac{1-r}{2} \sum_{j=1}^{p} \beta_j^2 \right) \tag{5.4}$$

where $l(\theta)$ is the negative log of partial likelihood described in Eq. (5.1) and α is a hyper-parameter that controls the amount of shrinkage and r is the relative weights of Lasso and Ridge penalties. we set r to 0.9 and for choosing the optimum value of α, 10 fold cross-validation is used on the training set of each initial folds. Initial folds are the folds that we used for MIL. Importance of features based on their coefficients in the best model are shown in Fig. 5.5 . Each fold indicate different ordering for feature importance. However some features that have non-zero coefficients are common in all three folds. Feature 12, 38 and 15 are the common features. Feature 12 belongs to the statistics of nuclear bounding box area and it indicates the standard deviations of the bounding box area. Feature 15 and feature 38 are minimum of nuclei eccentricity values and maximum of nuclei orientation degrees, respectively.

5.7.1 Patient Stratification

For computing the final score - -RPM score- - and stratifying patients into the low risk and high risk, the Cox models mentioned in the previous section are used. More precisely, for each fold, best hyper-parameter and coefficients are selected based on cross-validation, and the risk score is calculated. The best threshold for stratifying patients into low risk and high risk is selected based on the training sets. The cut-off threshold is a value that may best differentiate between the survival probability of low and high risk patients.

For each fold, this threshold is computed separately and applied on the corresponding test set. Then the values of RPM score for each test set are

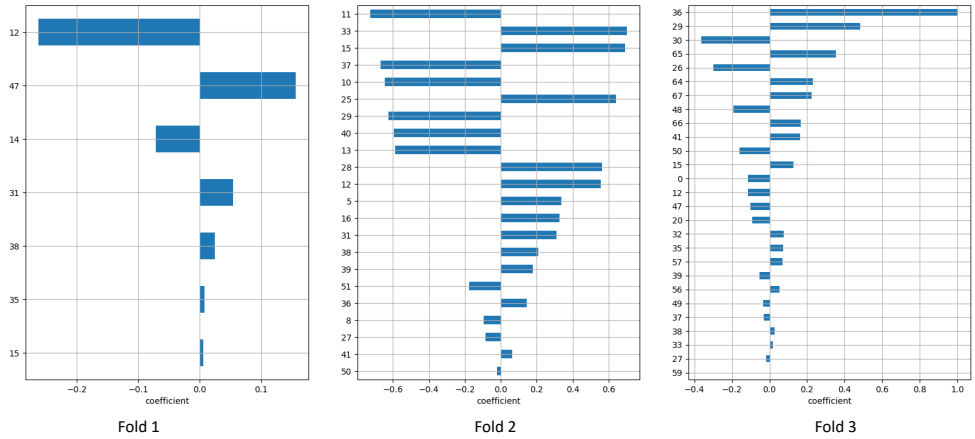

Figure 5.5: Model coefficients, value of each feature coefficient are shown which is indicative of feature importance. As initially we considered 3 folds, here 3 models are considered for obtaining best hyper parameters and coefficients.

Table 5.3: Univariate analysis for different feature, p-values using the log-rank test are reported.

Clinical Feature	β	HR (95% CI fo HR)	p-value
cancer type	-0.239	0.787 (0.583-1.063)	0.11
gender	-0.053	0.948 (0.701-1.283)	0.731
age	-0.009	0.991 (0.976-1.007)	0.265
stage	0.532	1.702 (1.461-1.983)	8.77e-12
model score (RPM)	0.949	2.584 (2.158-3.094)	<2e-16

concatenated to each other to form the RPM score for whole dataset.

5.7.2 Univariate Analysis

We explore the prognostic significance of each predictor independent to others. Kaplan Meier curve is used to visualize the difference between the survival probability for each predictor. Fig. 5.6 presents the survival curve along with log-rank test based p-values for disease specific survival of TCGA cohort. Kaplan Meier curve for stage as clinical parameter shows that it is associated with disease-free survival of lung patients ($p < 0.0001$). However, there is no association between survival and other clinical parameters like cancer type, age and gender. Kaplan-Meier curve for RPM score shows a clear separation between low and high risk patients and proves its prognostic value based on p-value ($p < 0.0001$). Table Table 5.3 shows the univariate Cox analysis for the RPM model scoring, which has a significant p-value=0.0003. It shows that at a given time instant, an increment of one unit for RPM score increase the risk of dying 2.584 times more.

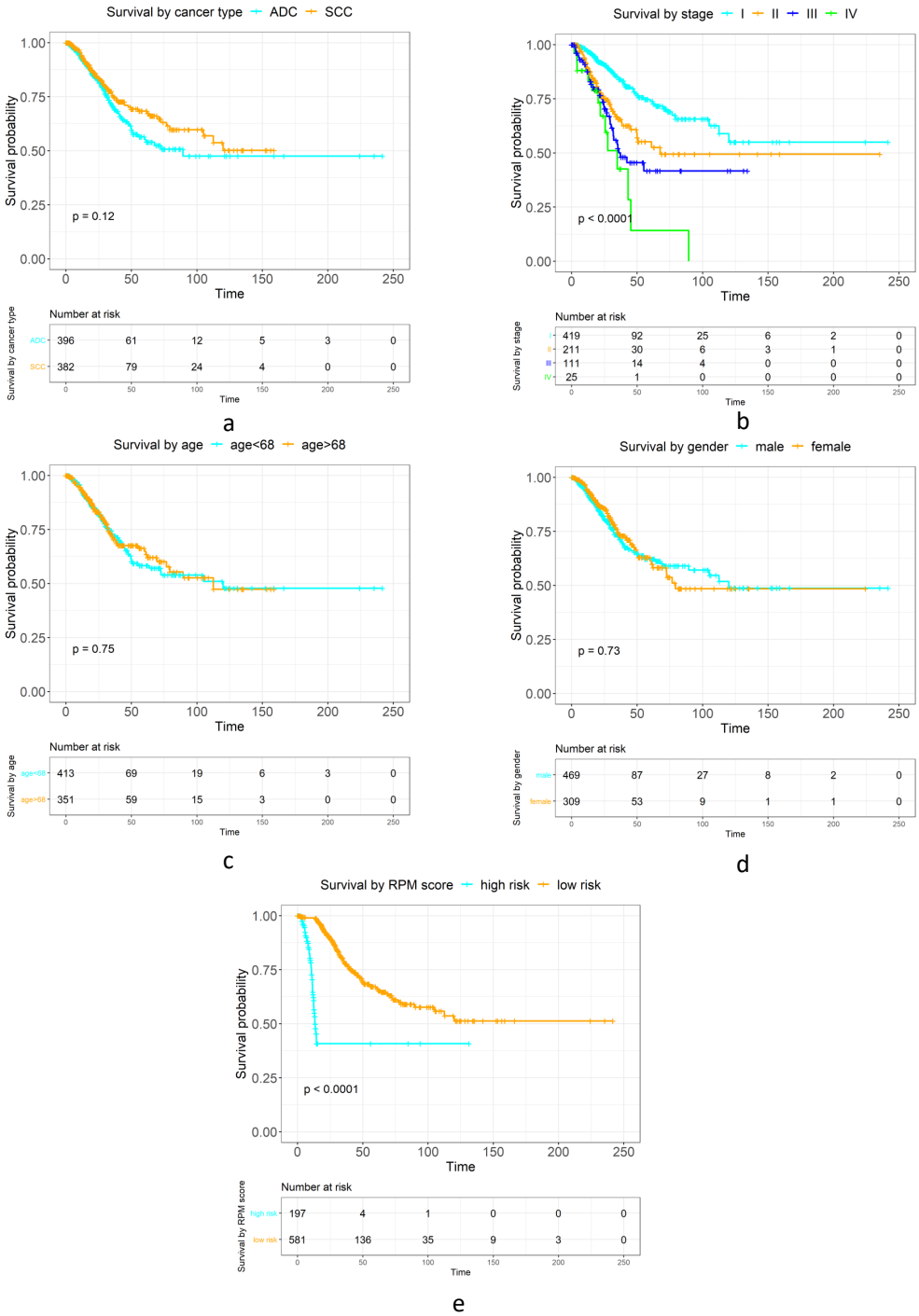

Figure 5.6: Kapalan Meier curves along with log-rank test based p-values for disease specific survival using different variables. a-d Kaplan Meier curves for variable available in TCGA and e. Kaplan Meier curve for the score that we obtained using our pipeline.

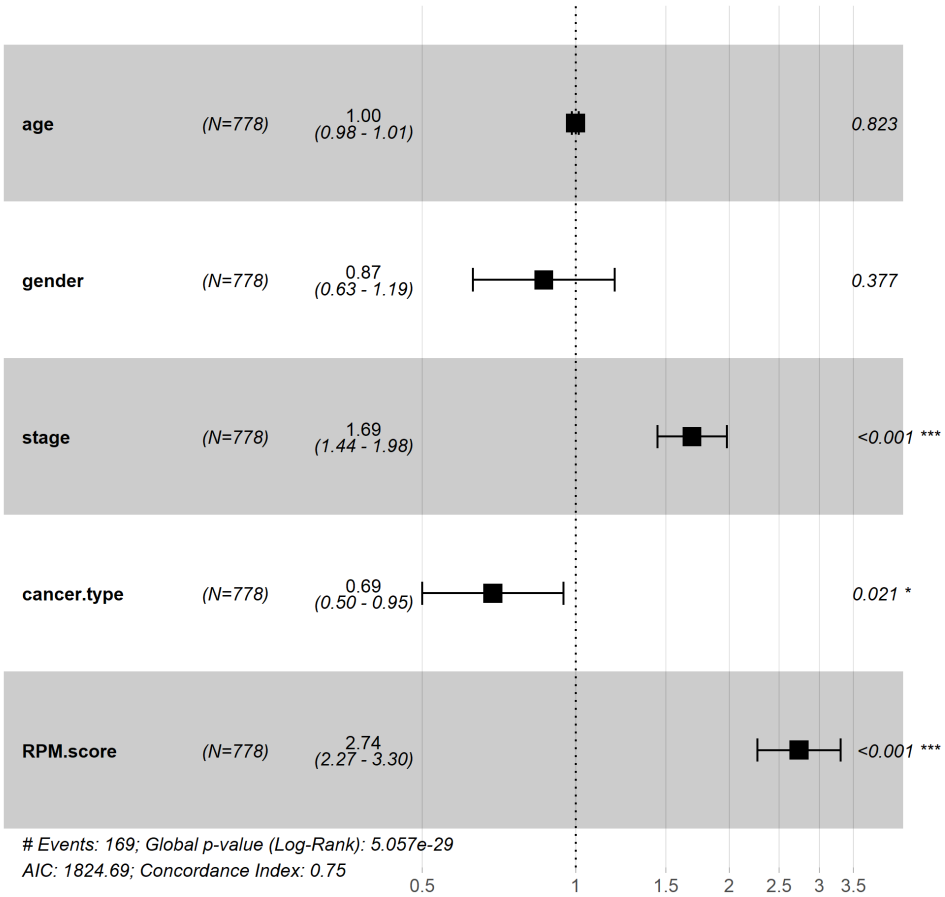

Figure 5.7: Multivariate analysis of the RPM-scoring in the presence of available variables in lung TCGA cohort for disease specific survival

5.7.3 Multivariate Analysis

Fig. 5.7 shows the multivariate analysis using the proposed RPM scoring and the clinicopathological variables whose information is available for lung TCGA cohort for disease specific survival. The score obtained from our pipeline is significant indicator of patient survival ($p < 0001$, HR=2.74 95%CI 2.27-3.30). It means that at a given time, for a patient with an increment of one unit for RPM score, the risk of dying is 2.584 more than a patient with one unit less RPM score. Similar to uni-variate analysis, cancer stage is also prognostic ($p < 0.001$, HR=1.69, 95%CI 1.44-1.98) in the presence of other variables. Age and gender do not show any significance in our analysis. However, unlike univariate analysis, cancer type is also a prognostic factor adjusting for other variables ($p = 0.021$ HR=0.69, 95%CI 0.50-0.95). Overall, these results show that the morphological features extracted from representative patches could be used in a Cox model for predicting a prognostic score.

5.8 Summary

In this chapter, we presented a pipeline for extracting features from representative patches of WSIs, without accessing tumour annotations. We have used attention-based multiple instance learning to select the most representative patches based on their attention values. Multiple instance learning model was trained with different target labels. We observed that in our experimental setup, distinguishing between the NSCLC subtype have the best performance. Therefore this model was chosen for patch selection.

We extracted 68 different morphological features from top 100 patches of each WSI. And their average values were used to optimize the coefficients of Cox proportional hazard model. We showed that three features are potentially highly correlated with patient survival. Moreover, the RPM scoring (the risk score obtained from Cox model) is a prognostic score in both univariate and multivariate analysis.

Chapter 6

Conclusions and Future Directions

In this chapter, we summarize the methods presented in this thesis and discuss some of the possible future directions for further exploration of the concepts.

In this thesis, we proposed a set of automated methods for analysis of histology images aimed to address some of the major challenges in computational pathology. We presented methods for segmentation and detection of nuclei in histology images where these methods were used for extracting a set of morphological features. We showed that these features are prognostically significant and are worth further exploring for their clinical significance. Moreover, throughout this thesis, we proposed a set of methods for utilizing minimum annotation for classification of histology images.

The proposed methods include: 1) an algorithm for nuclear detection, 2) an algorithm for instance segmentation of nuclei in histology images, 3) an interactive method for gland and nuclear segmentation, 4) domain-specific self-supervision approaches to deal with limited budget of annotations in histology, and 5) a framework based on multiple instance learning to extract highly attended patches where these patches are used for morphological analysis.

Throughout this thesis, we have conducted experiments on both image patches and WSIs. The first 3 chapters are mainly concerned with localization of nuclei in histology images. The localization approaches are mainly the basis for further analysis of nuclear morphometry. Although deep learning end-to-end approaches showed promising results in various tasks such as regression, classification, detection, predicting survival and etc, their explainability is limited and in most cases it is hard to interpret their decision. Therefore, investigation of the features extracted from objects such as nuclei and glands can shed some light on the underlying behaviour of the tumour micro-environment. One main challenge in histology image analysis is the scarcity of labels or weakly annotated dataset. In this thesis, we tried to overcome these challenges from

two perspectives: 1) Develop a method that can generate robust and trustable annotations 2) Develop a method that utilize limited budget of annotations while having good performance.

Below we summarize each of the methods presented in this thesis and we discuss potential future directions.

6.1 Mixture Density Networks for Nuclear Detection

In Chapter 2, we showed an application of mixture density networks for nuclear detection. Mixture density networks are suitable to map a single input to several possible outputs and we utilize this property to detect multiple nuclei in a single image patch. A new modified form of a cost function based on mixture densities is proposed for training the deep model. We have used the mixture of Gaussian distributions where the centroids of nuclei are the means of Gaussian distributions. The model predicts the centroids and the uncertainty of nuclear locations where the local maxima is used to detect the final location of nuclei. This approach might not perform well when the number of nuclei in a single patch increases. Since the loss function consists of multiple distributions, all densities might converge to a single point. One remedy would be decreasing the number of mixtures in the loss function. To do so, we can split the image into smaller grids and then apply loss for each region separately. This idea would be similar to how YOLO algorithm formulates the problem but in the context of mixture density networks.

Another possible future direction is applying Expectation Maximization (EM) algorithm after we find the nuclear location. This can potentially give better localization. To this end, the initial cluster centroids can be set by using mixture density networks and EM is used to refine the locations and even segmenting the nuclei.

6.2 Nuclear Instance Segmentation

We proposed a method for nuclear instance segmentation. We have used multi-scale blocks in our model to capture information at different scales. Positional information are fed into the model at different layers for predicting better segmentation results. In our pipeline, first the pixel-wise segmentation and centroid detection maps of nuclei are predicted with the dual-head variation of our proposed network. Afterwards, based on these outputs, a spatial information related to each nucleus instance is predicted using single head model. To separate the nuclei in ground truth, we replace pixels belonging to each nucleus with its spatial information (x/y coordinates of centroids, top left

coordinates of bounding box, etc.). To construct the final output, we apply a clustering algorithm on the predicted spatial map. For future work, we will consider the classes of nuclei as well, to this end one head can be added to determine classes of nuclei. Moreover, one may add other components to the model for its better performance in an end-to-end manner. To this end, we can increase the depth of network and add auxiliary loss function like dice loss.

6.3 Interactive Segmentation of Glands and Nuclei

Deep learning based models as the best performing models require huge amount of labeled data for precise and reliable prediction. However, collecting labeled data is expensive because it necessarily involves expert knowledge. Perhaps this is best demonstrated by medical tasks in which labels are the result of a time-consuming analysis made by multiple human experts. As nuclei, cells and glands are fundamental objects for downstream analysis in histology, in this chapter, we proposed a simple CNN-based approach to speed up collecting annotations for these objects which requests for minimum interaction from the annotator. We showed that for nuclei and cells as small objects, one click inside each object is enough for NuClick to yield a precise annotation. For glands as large objects, we proposed a novel approach to provide NuClick with a squiggle as a guiding signal, enabling it to outline the exact gland boundaries. These supervisory signals are fed to the network as auxiliary inputs along with RGB channels. With detailed experiments, we show that NuClick is generalizable, robust against variations in the user input, adaptable to new domains, and delivers reliable annotations.

One possible extension of this work would be making the framework probabilistic. More precisely, we can learn a distribution over segmentations given an input. For example, a variational autoencoder can be utilized to produce several segmentation hypotheses. Therefore, when user clicks on the object, several segmentation hypotheses would be shown to him/her and then he/she selects the one that best fits the object of interest.

6.4 Self-supervision for Classification of Pathology Images with Limited Annotations

In Chapter 5, we investigated self-supervised pretext tasks for classification of histology images in the presence of limited budget of annotations. We proposed 3 domain-specific self-supervised tasks and we showed that they can improve performance of annotations when the budget of annotations is very low. These self-supervised tasks are trained simultaneously with the main task where their

backbone is shared, therefore the backbone learns the representations that generalize well.

This approach can be applied to other problems such as segmentation in computational pathology, where annotation budget is often limited or large amount of labeled image data is not available. Another future direction could be employing other self-supervision tasks such as predicting the Eosin channel or a combination of Hematoxylin and Eosin after estimating the two channels, rather than keeping them fixed, and increasing the JigMag grids to incorporate wider and complex puzzles for the network to solve.

6.5 Morphological Features of Representative Patches to Predict NSCLC Survival

Most approaches in survival analysis rely on extracting patches from ROIs delineated by experts. Annotating these ROIs is time consuming and infeasible when use WSIs at large scale. In this chapter we have utilized attention-based multiple instance learning, which takes all the patches from WSIs in the form of a bag and assign a label to this bag. Attention block in the model assigns more weight to the patches that contribute more to to determining the class of whole bag. We observe that determining the sub-class of lung images achieve better performance than predicting stage, survival. Therefore the MIL was trained using lung sub-types (ACC, SCC) and afterwards we select the top k patches with highest weights (which mostly belong to tumour class). Then morphological analysis was performed on these patches and it has been shown that the features obtained from these selective patches are predictive of survival.

One possible extension to this work can be extracting other types of features such as texture and contextual features and see their prognostic value in our setting. Moreover, for our future work we will conduct this approach on the external test cohort to assess the generalizibility of our algorithm.

6.6 Concluding Remarks

In this thesis, we have presented set of works ranging from localization of nuclei to predicting the survival using those localization techniques. Tackling the challenge of limited annotation budget is another path that has been explored where we proposed techniques to overcome the challenges related to scarcity of annotations.

The results presented in Chapter 5 are preliminary and requires more extensive experiments on external and large cohorts. Interpretibility and transparency of the algorithm's decision are two objectives that deep models still

struggle with. There is plenty of room to explore features that are understandable by humans for disease diagnosis and prognosis. Such features should be assessed by experts for their potential use in clinical practise. Therefore, detailed investigation into the work presented in Chapter 5 and other similar approaches is required to understand the tumour micro-enviroment by exploring the usefulness of handcrafted features for predicting patient outcome.

Appendix A

Self-Path

A.1 Network Architecture

The performance on classification tasks was evaluated using supervised learning. ResNet50 was chosen since it has overall good performance while having lower number of parameters. The AUC-ROC performances can be seen in Table A.1. ResNet50 was used as the backbone architecture in all the self-supervision experiments except when the generative real vs fake prediction had to be used. While using the real vs fake auxiliary task for image generation, we utilize the architecture presented in Table A.2 and find that this simpler feature extractor allows easy and robust convergence for the image generator.

A.2 Hyper-Parameters

The hyper-parameters when using the various network architectures for training are shown in Table A.4 and Table A.5. Table A.4 is the hyper-parameter setting when using ResNet50 as the backbone and Table A.5 are the settings used when the generative real vs fake sub-task is used.

Table A.1: Performance of different baseline models on the three datasets. The evaluation was done using only the supervised loss and keeping the labeling budget at one percent.

	Kather	Camleyon16	LNM-OSCC
Labeled patches	800	600	134
Resnet50	0.9137	0.6467	0.7387
Resnet101	0.9015	0.6515	0.7314
Densenet121	0.9014	0.6514	0.7265
InceptionV3	0.8914	0.6618	0.7264

Table A.2: Network architecture while using the generative real vs fake subtask. Conv.T stands for transposed convolution.

Generator
latent space (100)
dense 4×4×512 batchnorm ReLU
5×5 Conv.T 512 batchnorm ReLU stride=2
5×5 Conv.T 256 batchnorm ReLU stride=2
5×5 Conv.T 128 batchnorm ReLU stride=2
5×5 Conv.T 128 batchnorm ReLU stride=2
5×5 Conv.T 3 weightnorm Tanh stride=2
Discriminator
128×128×3 images
dropout, $p = 0.2$
3×3 conv. weightnorm 96 lReLU
3×3 conv. weightnorm 96 lReLU
3×3 conv. weightnorm lReLU stride=2
dropout, $p = 0.5$
3×3 conv. weightnorm 128 lReLU
3×3 conv. weightnorm 128 lReLU
3×3 conv. weightnorm 128 lReLU stride=2
dropout, $p = 0.5$
3×3 conv. weightnorm 192 lReLU
3×3 conv. weightnorm 192 lReLU
3×3 conv. weightnorm 192 lReLU stride=2
dropout, $p = 0.5$
3×3 conv. weightnorm 192 lReLU
3×3 conv. weightnorm 192 lReLU
3×3 conv. weightnorm 192 lReLU
Adaptive maxpool
weightnorm dense 2

Table A.3: Network architecture for hematoxylin/decoder tasks

Decoder
Resnet50 backbone
1×1 Conv.T 512 ReLU stride=1
BilinearUpsample scale_factor=2
3×3 Conv.T 512 ReLU stride=1
BilinearUpsample scale_factor=2
3×3 Conv.T 256 ReLU stride=1
BilinearUpsample scale_factor=2
3×3 Conv.T 256 ReLU stride=1
BilinearUpsample scale_factor=2
3×3 Conv.T 128 ReLU stride=1
BilinearUpsample scale_factor=2
3×3 Conv.T 65 ReLU stride=1
1×1 Conv.T **Number of classes** stride=1

Table A.4: Hyper-parameters of model when Resnet 50 is used as feature extractor

Hyperparameters	Values
Batch size	64
Epoch	200
Optimizer	ADAM ($\alpha = 3 * 10^{-3}$, $\beta_1 = 0.9$)

Table A.5: Hyper-parameters for real vs fake prediction subtask

Hyperparameters	Values
Batch size	32
Epoch	500
Leaky ReLU slope	0.2
Exp. moving average decay	0.999
Optimizer	ADAM ($\alpha = 3 * 10^{-4}$, $\beta_1 = 0.5$)
Weight initialization	Isotropic gaussian ($\mu = 0$, $\sigma = 0.05$)
Bias initialization	Constant (0)

Bibliography

[1] *What is cancer?*, 2020 (accessed December 19, 2020). https://www.cancer.org/cancer/cancer-basics/what-is-cancer.html.

[2] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[3] Robert Lanza, John Gearhart, Brigid Hogan, Douglas Melton, Roger Pedersen, E Donnall Thomas, James A Thomson, and Michael West. *Essentials of stem cell biology*. Elsevier, 2005.

[4] David Weller, Peter Vedsted, Greg Rubin, FM Walter, Jon Emery, S Scott, C Campbell, Rikke Sand Andersen, William Hamilton, Frede Olesen, et al. The aarhus statement: improving design and reporting of studies on early cancer diagnosis. *British journal of cancer*, 106(7):1262–1267, 2012.

[5] *What is metastatic cancer?*, 2020 (accessed November 20, 2020). https://www.cancer.org/treatment/understanding-your-diagnosis/advanced-cancer.

[6] Jane Lea, Gideon Bachar, Anna M Sawka, Deepak C Lakra, Ralph W Gilbert, Jonathan C Irish, Dale H Brown, Patrick J Gullane, and David P Goldstein. Metastases to level iib in squamous cell carcinoma of the oral cavity: a systematic review and meta-analysis. *Head & Neck: Journal for the Sciences and Specialties of the Head and Neck*, 32(2):184–190, 2010.

[7] Juliana Noguti and et al. Metastasis from oral cancer: an overview. *Cancer Genomics-Proteomics*, 9(5):329–335, 2012.

[8] *Lung Cancer*, 2020 (accessed November 21, 2020). https://www.nhs.uk/conditions/lung-cancer.

[9] Jon Zugazagoitia, Ana Belen Enguita, Juan Antonio Nuñez, Lara Iglesias, and Santiago Ponce. The new iaslc/ats/ers lung adenocarcinoma classification from a clinical perspective: current concepts and future prospects. *Journal of thoracic disease*, 6(Suppl 5):S526, 2014.

[10] Takashi Eguchi, Kyuichi Kadota, Bernard J Park, William D Travis, David R Jones, and Prasad S Adusumilli. The new iaslc-ats-ers lung adenocarcinoma classification: what the surgeon should know. In *Seminars in thoracic and cardiovascular surgery*, volume 26, pages 210–222. Elsevier, 2014.

[11] Bhagavathi Ramamurthy, Frederick D Coffman, and Stanley Cohen. A perspective on digital and computational pathology. *Journal of pathology informatics*, 6, 2015.

[12] Mike May. A better lens on disease. *Scientific American*, 302(5):74–77, 2010.

[13] Adam Goode, Benjamin Gilbert, Jan Harkes, Drazen Jukic, and Mahadev Satyanarayanan. Openslide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics*, 4, 2013.

[14] Adnan Mujahid Khan, Nasir Rajpoot, Darren Treanor, and Derek Magee. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering*, 61(6):1729–1738, 2014.

[15] Babak Ehteshami Bejnordi, Mitko Veta, and Van Diest. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

[16] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[17] Xue-Wen Chen and Xiaotong Lin. Big data deep learning: challenges and perspectives. *IEEE access*, 2:514–525, 2014.

[18] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[19] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[20] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[21] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

[22] Vicente Grau, AUJ Mewes, M Alcaniz, Ron Kikinis, and Simon K Warfield. Improved watershed transform for medical image segmentation using prior information. *IEEE transactions on medical imaging*, 23(4):447–458, 2004.

[23] George Lee, Robert W Veltri, Guangjing Zhu, Sahirzeeshan Ali, Jonathan I Epstein, and Anant Madabhushi. Nuclear shape and architecture in benign fields predict biochemical recurrence in prostate cancer patients following radical prostatectomy: preliminary findings. *European urology focus*, 3(4-5):457–466, 2017.

[24] Thanh Tran, Oh-Heum Kwon, Ki-Ryong Kwon, Suk-Hwan Lee, and Kyung-Won Kang. Blood cell images segmentation using deep learning semantic segmentation. In *2018 IEEE International Conference on Electronics and Communication Engineering (ICECE)*, pages 13–16. IEEE, 2018.

[25] Nicola Bougen-Zhukov, Sheng Yang Loh, Hwee Kuan Lee, and Lit-Hsin Loo. Large-scale image-based screening and profiling of cellular phenotypes. *Cytometry Part A*, 91(2):115–125, 2017.

[26] Pedro Quelhas, Monica Marcuzzo, Ana Maria Mendonça, and Aurélio Campilho. Cell nuclei and cytoplasm joint segmentation using the sliding band filter. *IEEE Transactions on Medical Imaging*, 29(8):1463–1473, 2010.

[27] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition, pages 3431–3440, 2015.

[30] Bahram Parvin, Qing Yang, Ju Han, Hang Chang, Bjorn Rydberg, and Mary Helen Barcellos-Hoff. Iterative voting for inference of structural saliency and characterization of subcellular events. *IEEE Transactions on Image Processing*, 16(3):615–623, 2007.

[31] Xin Qi, Fuyong Xing, David J Foran, and Lin Yang. Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set. *IEEE Transactions on Biomedical Engineering*, 59(3):754–765, 2012.

[32] Adel Hafiane, Filiz Bunyak, and Kannappan Palaniappan. Fuzzy clustering and active contours for histopathology image segmentation and nuclei detection. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 903–914. Springer, 2008.

[33] Lin Yang, Oncel Tuzel, Peter Meer, and David J Foran. Automatic image analysis of histopathology specimens using concave vertex graph. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 833–841. Springer, 2008.

[34] Hui Kong, Metin Gurcan, and Kamel Belkacem-Boussaid. Partitioning histopathological images: an integrated framework for supervised color-texture segmentation and cell splitting. *IEEE transactions on medical imaging*, 30(9):1661–1677, 2011.

[35] Hatice Cinar Akakin, Hui Kong, Camille Elkins, Jessica Hemminger, Barrie Miller, Jin Ming, Elizabeth Plocharczyk, Rachel Roth, Mitchell Weinberg, Rebecca Ziegler, et al. Automated detection of cells from immunohistochemically-stained tissues: application to ki-67 nuclei staining. In *Medical Imaging 2012: Computer-Aided Diagnosis*, volume 8315, page 831503. International Society for Optics and Photonics, 2012.

[36] Chanho Jung and Changick Kim. Segmenting clustered nuclei using h-minima transform-based marker extraction and contour parameterization. *IEEE transactions on biomedical engineering*, 57(10):2600–2604, 2010.

[37] Ke Zhi Mao, Peng Zhao, and Puay-Hoon Tan. Supervised learning-based cell image segmentation for p53 immunohistochemistry. *IEEE Transactions on Biomedical Engineering*, 53(6):1153–1163, 2006.

[38] Xiaodong Yang, Houqiang Li, and Xiaobo Zhou. Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and kalman

filter in time-lapse microscopy. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 53(11):2405–2414, 2006.

[39] Rintu Maria Thomas and Jisha John. A review on cell detection and segmentation in microscopic images. In *Circuit, Power and Computing Technologies (ICCPCT), 2017 International Conference on*, pages 1–5. IEEE, 2017.

[40] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 411–418. Springer, 2013.

[41] Yuanpu Xie, Fuyong Xing, Xiaoshuang Shi, Xiangfei Kong, Hai Su, and Lin Yang. Efficient and robust cell detection: A structured regression approach. *Medical image analysis*, 44:245–254, 2018.

[42] Jun Xu, Lei Xiang, Qingshan Liu, Hannah Gilmore, Jianzhong Wu, Jinghai Tang, and Anant Madabhushi. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE transactions on medical imaging*, 35(1):119–130, 2016.

[43] Hai Su, Fuyong Xing, Xiangfei Kong, Yuanpu Xie, Shaoting Zhang, and Lin Yang. Robust cell detection and segmentation in histopathological images using sparse reconstruction and stacked denoising autoencoders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 383–390. Springer, 2015.

[44] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206, 2016.

[45] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[46] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International*

Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

[48] Yanning Zhou, Omer Fahri Onder, Qi Dou, Efstratios Tsougenis, Hao Chen, and Pheng-Ann Heng. Cia-net: Robust nuclei instance segmentation with contour-aware information aggregation. In *International Conference on Information Processing in Medical Imaging*, pages 682–693. Springer, 2019.

[49] Hao Chen, Xiaojuan Qi, Lequan Yu, Qi Dou, Jing Qin, and Pheng-Ann Heng. Dcan: Deep contour-aware networks for object instance segmentation from histology images. *Medical image analysis*, 36:135–146, 2017.

[50] Hirohisa Oda, Holger R Roth, Kosuke Chiba, Jure Sokolić, Takayuki Kitasaka, Masahiro Oda, Akinari Hinoki, Hiroo Uchida, Julia A Schnabel, and Kensaku Mori. Besnet: boundary-enhanced segmentation of cells in histopathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 228–236. Springer, 2018.

[51] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE transactions on medical imaging*, 38(2):448–459, 2018.

[52] Quoc Dang Vu, Simon Graham, Tahsin Kurc, Minh Nguyen Nhat To, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Jayashree Kalpathy-Cramer, Tianhao Zhao, et al. Methods for segmentation and classification of digital microscopy tissue images. *Frontiers in bioengineering and biotechnology*, 7, 2019.

[53] Christopher M Bishop. Mixture density networks. Technical report, Citeseer, 1994.

[54] Weidi Xie, J Alison Noble, and Andrew Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization*, 6(3):283–292, 2018.

[55] Yuanpu Xie, Fuyong Xing, Xiangfei Kong, Hai Su, and Lin Yang. Beyond classification: structured regression for robust cell detection using convolutional neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 358–365. Springer, 2015.

[56] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[57] Manohar Kuse, Yi-Fang Wang, Vinay Kalasannavar, Michael Khan, and Nasir Rajpoot. Local isotropic phase symmetry measure for detection of beta cells and lymphocytes. *Journal of pathology informatics*, 2, 2011.

[58] Yinyin Yuan, Henrik Failmezger, Oscar M Rueda, H Raza Ali, Stefan Gräf, Suet-Feung Chin, Roland F Schwarz, Christina Curtis, Mark J Dunning, Helen Bardwell, et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Science translational medicine*, 4(157):157ra143–157ra143, 2012.

[59] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[60] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[61] Mostafa Jahanifar, Neda Zamani Tajeddin, Navid Alemi Koohbanani, Ali Gooya, and Nasir Rajpoot. Segmentation of skin lesions and their attributes using multi-scale convolutional neural networks and domain specific augmentations. *arXiv preprint arXiv:1809.10243*, 2018.

[62] Navid Alemi Koohababni, Mostafa Jahanifar, Ali Gooya, and Nasir Rajpoot. Nuclei detection using mixture density networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 241–248. Springer, 2018.

[63] Xiaodan Liang, Liang Lin, Yunchao Wei, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Proposal-free network for instance-level object segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2978–2991, 2017.

[64] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, 2017.

[65] Pavel Izmailov and et al. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[66] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[67] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[68] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[69] Saed Asgari Taghanaki, Kumar Abhishek, Joseph Paul Cohen, Julien Cohen-Adad, and Ghassan Hamarneh. Deep semantic segmentation of natural and medical images: A review. *accepted to appear in Springer Artificial Intelligence Review*, 2020.

[70] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

[71] Neeraj Kumar, Ruchika Verma, Deepak Anand, Yanning Zhou, Omer Fahri Onder, Efstratios Tsougenis, Hao Chen, Pheng Ann Heng, Jiahui Li, Zhiqiang Hu, et al. A multi-organ nucleus segmentation challenge. *IEEE transactions on medical imaging*, 2019.

[72] Inwan Yoo, Donggeun Yoo, and Kyunghyun Paeng. Pseudoedgenet: Nuclei segmentation only with point annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 731–739. Springer, 2019.

[73] Hui Qu, Pengxiang Wu, Qiaoying Huang, Jingru Yi, Gregory M Riedlinger, Subhajyoti De, and Dimitris N Metaxas. Weakly supervised deep nuclei segmentation using points annotation in histopathology images. In *International Conference on Medical Imaging with Deep Learning*, pages 390–400, 2019.

[74] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*, 2014.

[75] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016.

[76] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015.

[77] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7268–7277, 2018.

[78] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017.

[79] Bin Jin, Maria V Ortiz Segovia, and Sabine Susstrunk. Webly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3635, 2017.

[80] Ejaz Ahmed, Scott Cohen, and Brian Price. Semantic object selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3157, 2014.

[81] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016.

[82] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015.

[83] Ding-Jie Chen, Jui-Ting Chien, Hwann-Tzong Chen, and Long-Wen Chang. Tap and shoot segmentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[84] Tinghuai Wang, Bo Han, and John Collomosse. Touchcut: Fast image and video segmentation using single-touch interaction. *Computer Vision and Image Understanding*, 120:14–30, 2014.

[85] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3159–3167, 2016.

[86] Jia Xu, Alexander G Schwing, and Raquel Urtasun. Learning to segment under various forms of weak supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3781–3790, 2015.

[87] Xue Bai and Guillermo Sapiro. Geodesic matting: A framework for fast interactive image and video segmentation and matting. *International journal of computer vision*, 82(2):113–132, 2009.

[88] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. Interactively co-segmentating topically related images with intelligent scribble guidance. *International journal of computer vision*, 93(3):273–292, 2011.

[89] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 1, pages 105–112. IEEE, 2001.

[90] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[91] Ming-Ming Cheng, Victor Adrian Prisacariu, Shuai Zheng, Philip HS Torr, and Carsten Rother. Densecut: Densely connected crfs for realtime grabcut. In *Computer Graphics Forum*, volume 34, pages 193–201. Wiley Online Library, 2015.

[92] Varun Gulshan, Carsten Rother, Antonio Criminisi, Andrew Blake, and Andrew Zisserman. Geodesic star convexity for interactive image segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3129–3136. IEEE, 2010.

[93] Naveen Shankar Nagaraja, Frank R Schmidt, and Thomas Brox. Video segmentation with just a few strokes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3235–3243, 2015.

[94] Eric N Mortensen and William A Barrett. Interactive segmentation with intelligent scissors. *Graphical models and image processing*, 60(5):349–384, 1998.

[95] Stefano Cagnoni, Andrew B Dobrzeniecki, Riccardo Poli, and Jacquelyn C Yanch. Genetic algorithm-based interactive segmentation of 3d medical images. *Image and Vision Computing*, 17(12):881–895, 1999.

[96] Marleen de Bruijne, Bram van Ginneken, Max A Viergever, and Wiro J Niessen. Interactive segmentation of abdominal aortic aneurysms in cta images. *Medical Image Analysis*, 8(2):127–138, 2004.

[97] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7):1562–1573, 2018.

[98] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 577–585, 2018.

[99] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4930–4939, 2017.

[100] Vivek Kwatra, Arno Schödl, Irfan Essa, Greg Turk, and Aaron Bobick. Graphcut textures: image and video synthesis using graph cuts. In *ACM Transactions on Graphics (ToG)*, volume 22, pages 277–286. ACM, 2003.

[101] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep grabcut for object selection. *arXiv preprint arXiv:1707.00243*, 2017.

[102] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 373–381, 2016.

[103] Eirikur Agustsson, Jasper RR Uijlings, and Vittorio Ferrari. Interactive full image segmentation by considering all regions jointly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11622–11631, 2019.

[104] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 616–625, 2018.

[105] Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, and Sanja Fidler. Fast interactive object annotation with curve-gcn. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 5257–5266, 2019.

[106] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a polygon-rnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5230–5238, 2017.

[107] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 859–868, 2018.

[108] Zian Wang, David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Object instance annotation with deep extreme level set evolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7500–7508, 2019.

[109] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *International journal of computer vision*, 22(1):61–79, 1997.

[110] David Acuna, Amlan Kar, and Sanja Fidler. Devil is in the edges: Learning semantic boundaries from noisy annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11075–11083, 2019.

[111] Tomas Sakinis, Fausto Milletari, Holger Roth, Panagiotis Korfiatis, Petro Kostandy, Kenneth Philbrick, Zeynettin Akkus, Ziyue Xu, Daguang Xu, and Bradley J Erickson. Interactive segmentation of medical images through fully convolutional neural networks. *arXiv preprint arXiv:1903.08205*, 2019.

[112] Mykhaylo Andriluka, Jasper RR Uijlings, and Vittorio Ferrari. Fluid annotation: a human-machine collaboration interface for full image annotation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1957–1966, 2018.

[113] Claudia Nieuwenhuis and Daniel Cremers. Spatially varying color distributions for interactive multilabel segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1234–1247, 2012.

[114] Claudia Nieuwenhuis, Simon Hawe, Martin Kleinsteuber, and Daniel Cremers. Co-sparse textural similarity for interactive segmentation. In *European conference on computer vision*, pages 285–301. Springer, 2014.

[115] Jakob Santner, Thomas Pock, and Horst Bischof. Interactive multi-label segmentation. In *Asian Conference on Computer Vision*, pages 397–410. Springer, 2010.

[116] Vladimir Vezhnevets and Vadim Konouchine. Growcut: Interactive multi-label nd image segmentation by cellular automata. In *proc. of Graphicon*, volume 1, pages 150–156. Citeseer, 2005.

[117] Mostafa Jahanifar, Navid Alemi Koohbanani, and Nasir Rajpoot. Nuclick: From clicks in the nuclei to nuclear boundaries. *arXiv preprint arXiv:1909.03253*, 2019.

[118] Jiajun Wu, Yibiao Zhao, Jun-Yan Zhu, Siwei Luo, and Zhuowen Tu. Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 256–263, 2014.

[119] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, (23):3, 2012.

[120] Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019.

[121] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017.

[122] Jean Serra. *Image analysis and mathematical morphology.* Academic Press, Inc., 1983.

[123] Ruqayya Awan, Korsuk Sirinukunwattana, David Epstein, Samuel Jefferyes, Uvais Qidwai, Zia Aftab, Imaad Mujeeb, David Snead, and Nasir Rajpoot. Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. *Scientific reports*, 7(1):16852, 2017.

[124] Simon Graham, Hao Chen, Jevgenij Gamper, Qi Dou, Pheng-Ann Heng, David Snead, Yee Wah Tsang, and Nasir Rajpoot. Mild-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *Medical image analysis*, 52:199–211, 2019.

[125] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9404–9413, 2019.

[126] Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, 16(6):641–647, 1994.

[127] Tony F Chan and Luminita A Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001.

[128] K Parvati, Prakasa Rao, and M Mariya Das. Image segmentation using gray-scale morphology and marker-controlled watershed transformation. *Discrete Dynamics in Nature and Society*, 2008, 2008.

[129] Zaneta Swiderska-Chadaj, Hans Pinckaers, Mart van Rijthoven, Maschenka Balkenhol, Margarita Melnikova, Oscar Geessink, Quirine Manson, Mark Sherman, Antonio Polonia, Jeremy Parry, Mustapha Abubakar, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Learning to detect lymphocytes in immunohistochemistry with deep learning. *Medical Image Analysis*, 58:101547, 2019.

[130] Navid Alemi Koohbanani, Mostafa Jahanifar, Ali Gooya, and Nasir Rajpoot. Nuclear instance segmentation using a proposal-free spatially aware deep learning framework. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 622–630. Springer, 2019.

[131] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 2020.

[132] G Litjens. Automated slide analysis platform (asap). `http://rse.diagnijmegen.nl/software/asap/`, 2017.

[133] Peter Bankhead, Maurice B Loughrey, José A Fernández, Yvonne Dombrowski, Darragh G McArt, Philip D Dunne, Stephen McQuaid, Ronan T Gray, Liam J Murray, Helen G Coleman, et al. Qupath: Open source software for digital pathology image analysis. *Scientific reports*, 7(1):1–7, 2017.

[134] Ole-Johan Skrede, Sepp De Raedt, Andreas Kleppe, Tarjei S Hveem, Knut Liestøl, John Maddison, Hanne A Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, et al. Deep learning for prediction

of colorectal cancer outcome: a discovery and validation study. *The Lancet*, 395(10221):350–360, 2020.

[135] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[136] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

[137] Mohammad Peikari, Sherine Salama, Sharon Nofech-Mozes, and Anne L Martel. A cluster-then-label semi-supervised learning approach for pathology image classification. *Scientific reports*, 8(1):1–13, 2018.

[138] Jiayun Li, William Speier, King Chung Ho, Karthik V Sarma, Arkadiusz Gertych, Beatrice S Knudsen, and Corey W Arnold. An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies. *Computerized Medical Imaging and Graphics*, 69:125–133, 2018.

[139] Rachel Sparks and Anant Madabhushi. Out-of-sample extrapolation utilizing semi-supervised manifold learning (ose-ssl): content based image retrieval for histopathology images. *Scientific reports*, 6:27306, 2016.

[140] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[141] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[142] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4L: Self-supervised semi-supervised learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1476–1485, 2019.

[143] Simon Graham, David Epstein, and Nasir Rajpoot. Dense steerable filter cnns for exploiting rotational symmetry in histology images. *arXiv preprint arXiv:2004.03037*, 2020.

[144] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

[145] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.

[146] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019.

[147] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[148] Amit Kumar Jaiswal and et al. Semi-supervised learning for cancer detection of lymph node metastases. *arXiv preprint arXiv:1906.09587*, 2019.

[149] Hai Su and et al. Local and global consistency regularized mean teacher for semi-supervised nuclei classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 559–567. Springer, 2019.

[150] Shayne Shaw, Maciej Pajak, Aneta Lisowska, Sotirios A Tsaftaris, and Alison Q O'Neil. Teacher-student chain for efficient semi-supervised histology image classification. *arXiv preprint arXiv:2003.08797*, 2020.

[151] Ming Y Lu and et al. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv preprint arXiv:1910.10825*, 2019.

[152] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[153] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[154] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[155] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition, pages 3722–3731, 2017.

[156] Jian Ren, Ilker Hacihaliloglu, Eric A Singer, David J Foran, and Xin Qi. Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 201–209. Springer, 2018.

[157] Fuyong Xing, Tell Bennett, and Debashis Ghosh. Adversarial domain adaptation and pseudo-labeling for cross-modality microscopy image quantification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 740–749. Springer, 2019.

[158] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. A closer look at domain shift for deep learning in histopathology. *arXiv preprint arXiv:1909.11575*, 2019.

[159] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[160] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[161] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017.

[162] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[163] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5898–5906, 2017.

[164] Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

[165] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014.

[166] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.

[167] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

[168] Jacob Gildenblat and Eldad Klaiman. Self-supervised similarity learning for digital pathology. *arXiv preprint arXiv:1905.08139*, 2019.

[169] Herbert Freeman and L Garder. Apictorial jigsaw puzzles: The computer solution of a problem in pattern recognition. *IEEE Transactions on Electronic Computers*, (2):118–127, 1964.

[170] Arnout C Ruifrok, Dennis A Johnston, et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299, 2001.

[171] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[172] Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7345–7354, 2020.

[173] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31:3235–3246, 2018.

[174] Lei Cui, Hansheng Li, Wenli Hui, Sitong Chen, Lin Yang, Yuxin Kang, Qirong Bo, and Jun Feng. A deep learning-based framework for lung cancer survival analysis with biomarker interpretation. *BMC bioinformatics*, 21(1):1–14, 2020.

[175] Ulkü Yilmaz, Özlem Özmen, Funda Demirağ, Tuba İnal Cengiz, Pınar Akın Kabalak, Derya Kizilgöz, İbrahim Onur Alıcı, Göktürk Fındık,

et al. The relationship between quantitative positron emission tomography parameters, the invasive lung adenocarcinoma grading system of international association for the study of lung cancer/american thoracic society/european respiratory society, and survival. *Eurasian Journal of Pulmonology*, 21(2):107, 2019.

[176] Hongyuan Wang, Fuyong Xing, Hai Su, Arnold Stromberg, and Lin Yang. Novel image markers for non-small cell lung cancer classification and survival prediction. *BMC bioinformatics*, 15(1):310, 2014.

[177] Jiawen Yao, Sheng Wang, Xinliang Zhu, and Junzhou Huang. Imaging biomarker discovery for lung cancer survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 649–657. Springer, 2016.

[178] Kun-Hsing Yu, Ce Zhang, Gerald J Berry, Russ B Altman, Christopher Ré, Daniel L Rubin, and Michael Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications*, 7(1):1–10, 2016.

[179] Lee Kamentsky, Thouis R Jones, Adam Fraser, Mark-Anthony Bray, David J Logan, Katherine L Madden, Vebjorn Ljosa, Curtis Rueden, Kevin W Eliceiri, and Anne E Carpenter. Improved structure, function and compatibility for cellprofiler: modular high-throughput image analysis software. *Bioinformatics*, 27(8):1179–1180, 2011.

[180] Pranjal Vaidya, Xiangxue Wang, Kaustav Bera, Arjun Khunger, Humberto Choi, Pradnya Patil, Vamsidhar Velcheti, and Anant Madabhushi. Raptomics: integrating radiomic and pathomic features for predicting recurrence in early stage lung cancer. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 105810M. International Society for Optics and Photonics, 2018.

[181] Jun Cheng, Xiaokui Mo, Xusheng Wang, Anil Parwani, Qianjin Feng, and Kun Huang. Identification of topological features in renal tumor microenvironment associated with patient survival. *Bioinformatics*, 34(6):1024–1030, 2018.

[182] Knut Liestbl, Per Kragh Andersen, and Ulrich Andersen. Survival analysis and neural nets. *Statistics in medicine*, 13(12):1189–1200, 1994.

[183] W Nick Street. A neural network model for prognostic prediction. In *ICML*, pages 540–546, 1998.

[184] Leonardo Franco, José M Jerez, and Emilio Alba. Artificial neural networks and prognosis in medicine. survival analysis in breast cancer patients. In *ESANN*, pages 91–102, 2005.

[185] Elia Biganzoli, Patrizia Boracchi, Luigi Mariani, and Ettore Marubini. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, 17(10):1169–1186, 1998.

[186] David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

[187] Safoora Yousefi, Congzheng Song, Nelson Nauata, and Lee Cooper. Learning genomic representations to predict clinical outcomes in cancer. *arXiv preprint arXiv:1609.08663*, 2016.

[188] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

[189] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7234–7242, 2017.

[190] Jiawen Yao, Xinliang Zhu, Feiyun Zhu, and Junzhou Huang. Deep correlational learning for survival prediction from multi-modality data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 406–414. Springer, 2017.

[191] Jiawen Yao, Xinliang Zhu, and Junzhou Huang. Deep multi-instance learning for survival prediction from whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 496–504. Springer, 2019.

[192] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

[193] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.

[194] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[195] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[196] Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European Congress on Digital Pathology*, pages 11–19. Springer, 2019.

[197] Jevgenij Gamper, Navid Alemi Koohbanani, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. Pannuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020.

[198] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

[199] Shan E Ahmed Raza, Linda Cheung, Muhammad Shaban, Simon Graham, David Epstein, Stella Pelengaris, Michael Khan, and Nasir M Rajpoot. Micro-net: A unified model for segmentation of various objects in microscopy images. *Medical image analysis*, 52:160–173, 2019.